Goa Institute of Management

HR ANAYTICS

End Term Project

2021 – 2023

Term 6

Submitted by: Group 3 (Section B)

Group Details:

Group Members	Roll Numbers
Abeer Pareek	B2021054
Arpit Sharma	B2021065
Chavi Garg	B2021069
Jash Patel	B2021080
Keerthana K	B2021081
Sakshi Agarwal	B2021091
Subhesh Jha	B2021099
Sunaina M Bhagat	B2021101

Contents

1 Objective

To apply different machine learning models and check which can accurately predict which hospital employees are at risk of leaving their job, and to use this information to implement targeted interventions to reduce attrition rates and retain valuable staff members.

2 Hypothesis

We aim to discover if there is a significant difference in the attrition rate across different departments. We want to test if working overtime has a significant impact on employee attrition. Our goal is to investigate if the distance an employee must travel to work is a significant factor in their likelihood of experiencing attrition.

We seek to determine if job involvement is a significant factor in employee attrition. We want to test if there is a significant difference in attrition rates among various job roles. Our aim is to investigate whether an employee's travel pattern has an impact on their likelihood of experiencing attrition. We want to test if any specific education level is significantly associated with employee attrition.

3 Dataset Description

This data set gives details about the employees of a hospital, including their various attributes along with attrition condition.

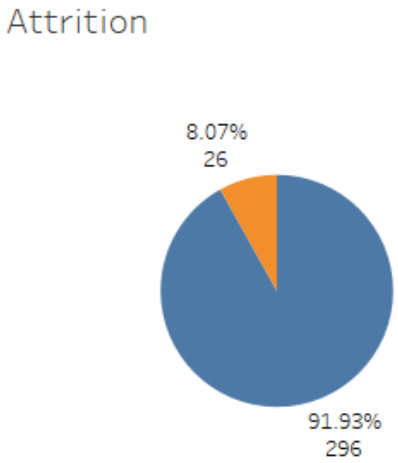
Employeel	Age	Attrition	BusinessTravel	DailyRate	Departmer	DistanceFr	Education	EducationField	Empl	Environmen	Gender	HourlyRate	JobInvolve	JobLevel	JobRole	JobSatisfac	MaritalStat	MonthlyIn
1379450	41	No	Travel_Rarely	465	Maternity	14	3	Life Sciences	1	1	Male	56	3	1	Other	3	Divorced	2451
1138977	30	No	Travel_Rarely	1339	Cardiology	5	3	Life Sciences	1	2	Female	41	3	3	Nurse	4	Married	9419
1726077	38	No	Travel_Rarely	702	Cardiology	1	4	Life Sciences	1	1	Female	59	2	2	Nurse	4	Single	8686
1101855	32	No	Travel_Rarely	120	Maternity	6	5	Life Sciences	1	3	Male	43	3	1	Other	3	Single	3038
1715001	27	No	Travel_Rarely	1157	Maternity	17	3	Technical Degree	1	3	Male	51	3	1	Other	2	Married	3058
1820041	19	Yes	Travel_Frequently	602	Cardiology	1	1	Technical Degree	1	3	Female	100	1	1	Other	1	Single	2325
1540435	36	No	Travel_Frequently	1480	Maternity	3	2	Medical	1	4	Male	30	3	1	Nurse	2	Single	2088
1340085	30	No	Non-Travel	111	Maternity	9	3	Medical	1	3	Male	66	3	2	Nurse	1	Divorced	3072
1319352	45	No	Travel_Rarely	1268	Cardiology	4	2	Life Sciences	1	3	Female	30	3	2	Nurse	1	Divorced	5006
1722423	56	No	Travel_Rarely	713	Maternity	8	3	Life Sciences	1	3	Female	67	3	1	Other	1	Divorced	4257
1853558	33	No	Travel_Rarely	134	Maternity	2	3	Life Sciences	1	3	Male	90	3	1	Other	4	Single	2500
1517594	19	Yes	Travel_Rarely	303	Maternity	2	3	Life Sciences	1	2	Male	47	2	1	Nurse	4	Single	1102
1700841	46	No	Travel_Rarely	526	Cardiology	1	2	Marketing	1	2	Female	92	3	3	Nurse	1	Divorced	10453
1665761	38	No	Travel_Rarely	1380	Maternity	9	2	Life Sciences	1	3	Female	75	3	1	Nurse	4	Single	2288
1224843	31	No	Travel_Rarely	140	Maternity	12	1	Medical	1	3	Female	95	3	1	Other	4	Married	3929
1739412	34	No	Travel_Rarely	629	Maternity	27	2	Medical	1	4	Female	95	3	1	Other	2	Single	2311
1190255	41	Yes	Travel_Rarely	1356	Cardiology	20	2	Marketing	1	2	Female	70	3	1	Other	2	Single	3140
1820323	50	No	Travel_Rarely	328	Maternity	1	3	Medical	1	3	Male	86	2	1	Nurse	3	Married	3690
1477195	53	No	Travel_Rarely	1084	Maternity	13	2	Medical	1	4	Female	57	4	2	Therapist	1	Divorced	4450

These variables contain details about the demographics (age, gender, education, marital status, etc.) of the employees along with their employment details such as monthly income, job role, department, performance rating.

4 EDA using Tableau

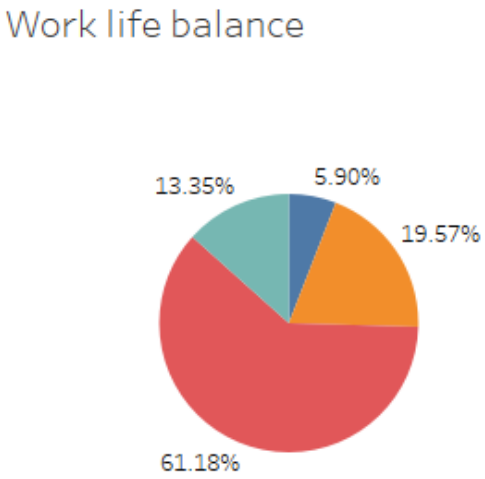
4.1 Univariate

4.1.1 Attrition:



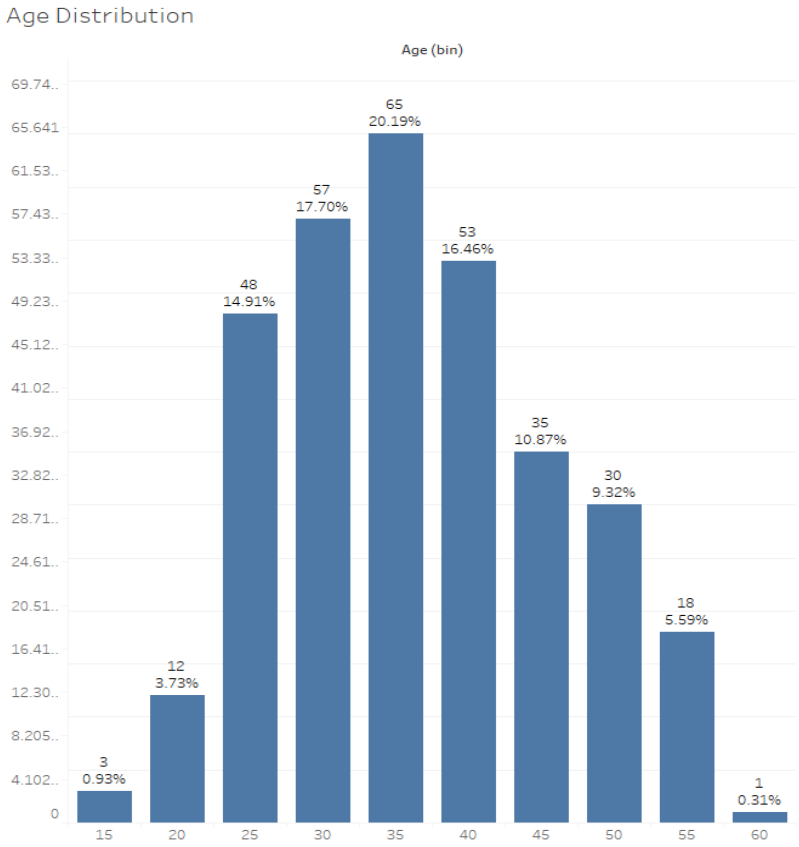
- Select the "Attrition" and "Employee Count" variables.
- Make a pie chart for the selected variables.
- Make a quick table calculation to find the percentage of employee attrition.
- Drag and drop "Emp Count" and "Emp Count Table Calculation" into the labels to display the count and percentage.

4.1.2 Work Life Balance:



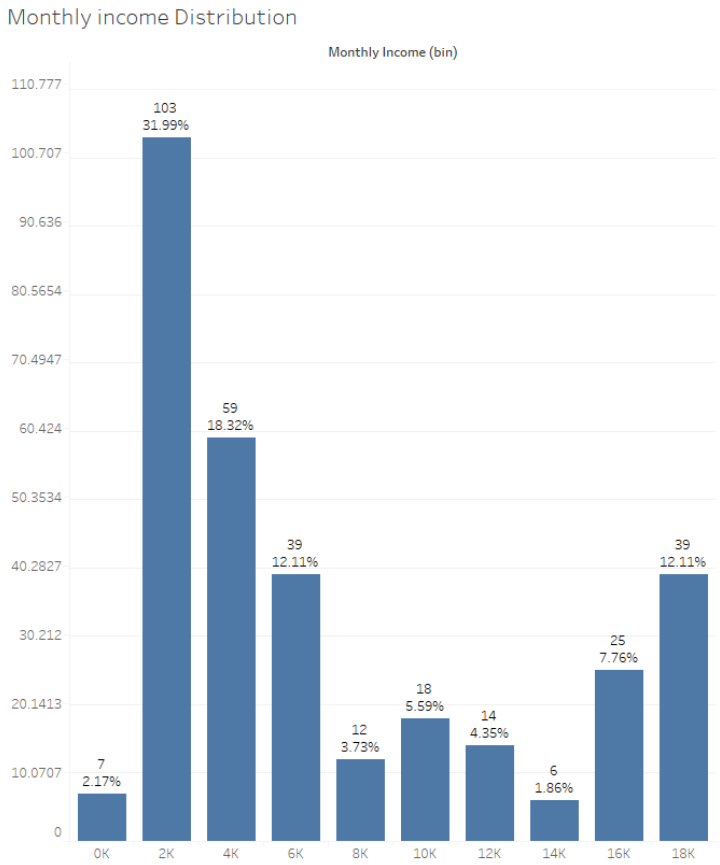
- Select the "Work Life Balance" and "Employee Count" variables.
- Change the "Work Life Balance" variable type from measure to dimension.
- Make a pie chart for the selected variables.
- Make a quick table calculation to find the percentage of employee attrition.
- Drag and drop "Emp Count" and "Emp Count Table Calculation" into the labels to display the count and percentage.

4.1.3 Age Distribution:



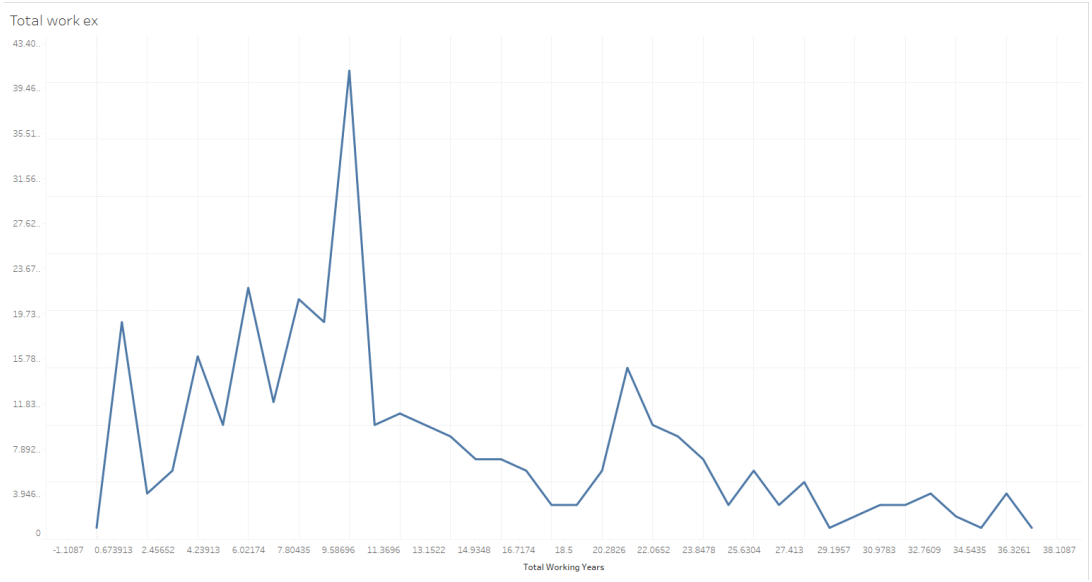
- Create a bin of age with bin value 5.
- Select "Age Bin" and "Emp Count".
- Make a quick table calculation to find the percentage of employee attrition.
- Drag and drop "Emp Count" and "Emp Count Table Calculation" into the labels to display the count and percentage.

4.1.4 Monthly income Distribution:



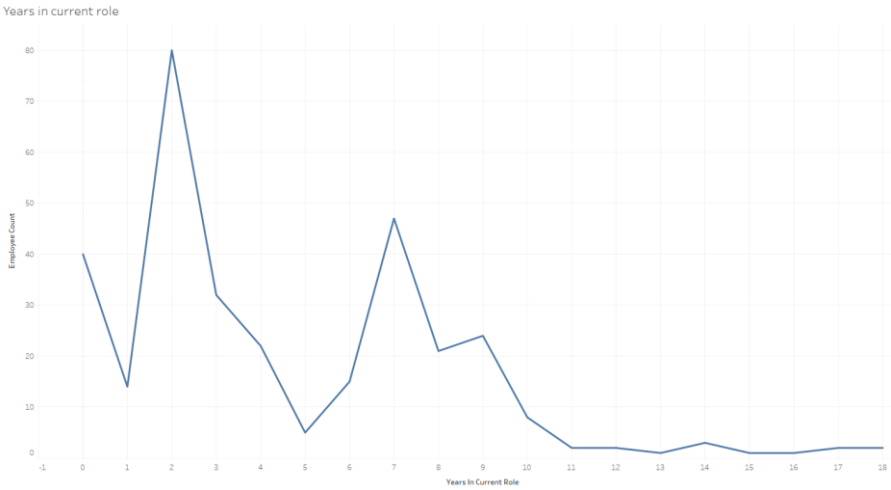
- Create a bin of monthly income with bin value 5.
- Select "Monthly Income Bin" and "Emp Count".
- Make a quick table calculation to find the percentage of employee attrition.
- Drag and drop "Emp Count" and "Emp Count Table Calculation" into the labels to display the count and percentage.

4.1.5 Total Work ex:



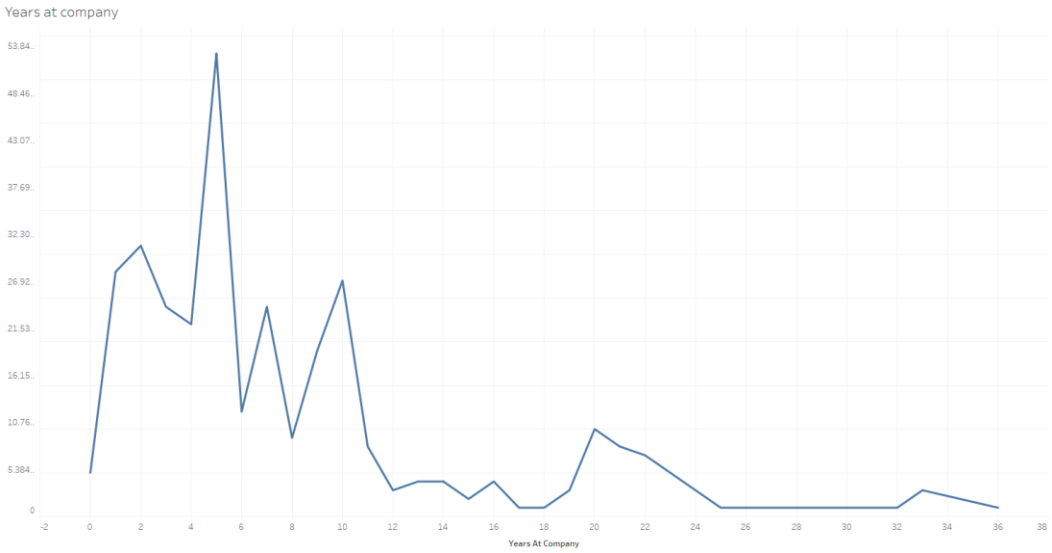
- Select "Total Work Ex" and "Emp Count".
- Change the data type of "Total Work Ex" from measure to dimension.
- Select a line chart to display the distribution.

4.1.6 Years in current role:



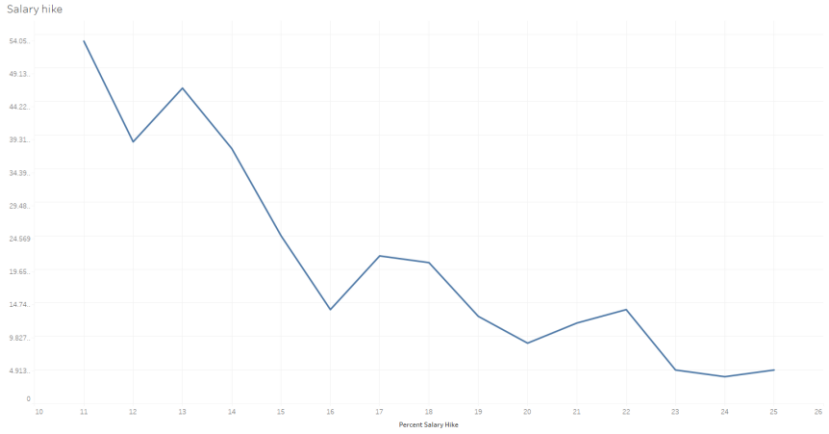
- Select "Years in Current Role" and "Emp Count".
- Change the data type of "Years in Current Role" from measure to dimension.
- Select a line chart to display the distribution.

4.1.7 Years in company:



- Select "Years at Company" and "Emp Count".
- Change the data type of "Years at Company" from measure to dimension.
- Select a line chart to display the distribution.

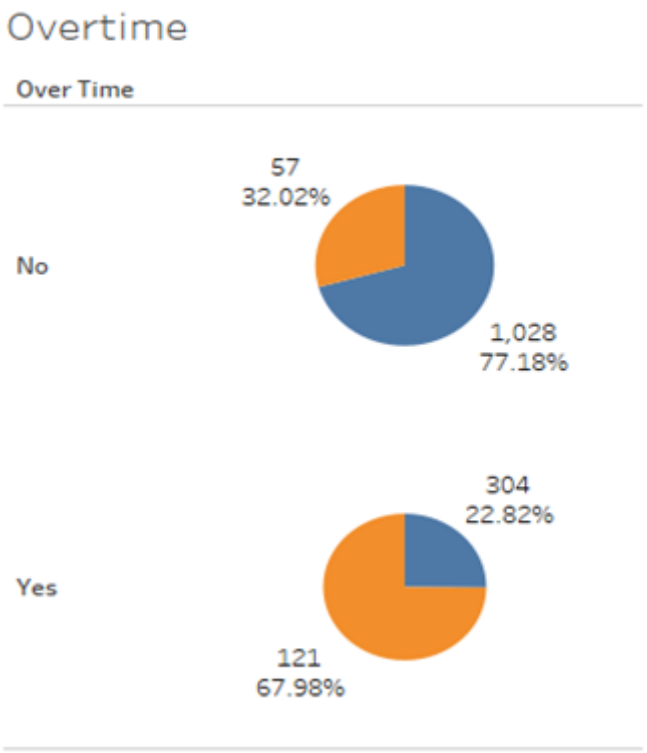
4.1.8 Salary hike:



- Select "Salary Hike" and "Emp Count".
- Change the data type of "Salary Hike" from measure to dimension.
- Select a line chart to display the distribution.

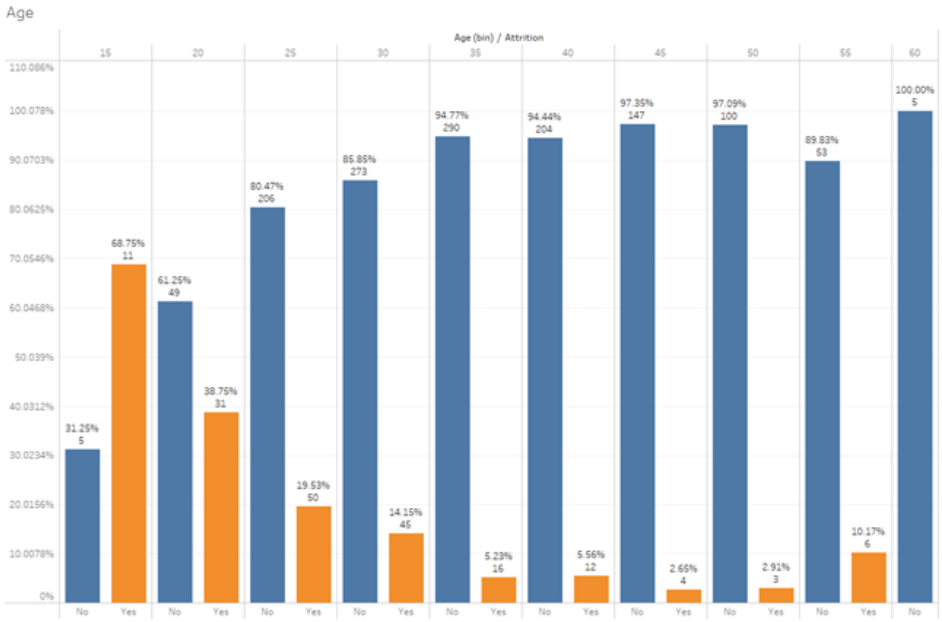
4.2 Bivariate

4.2.1 Overtime:



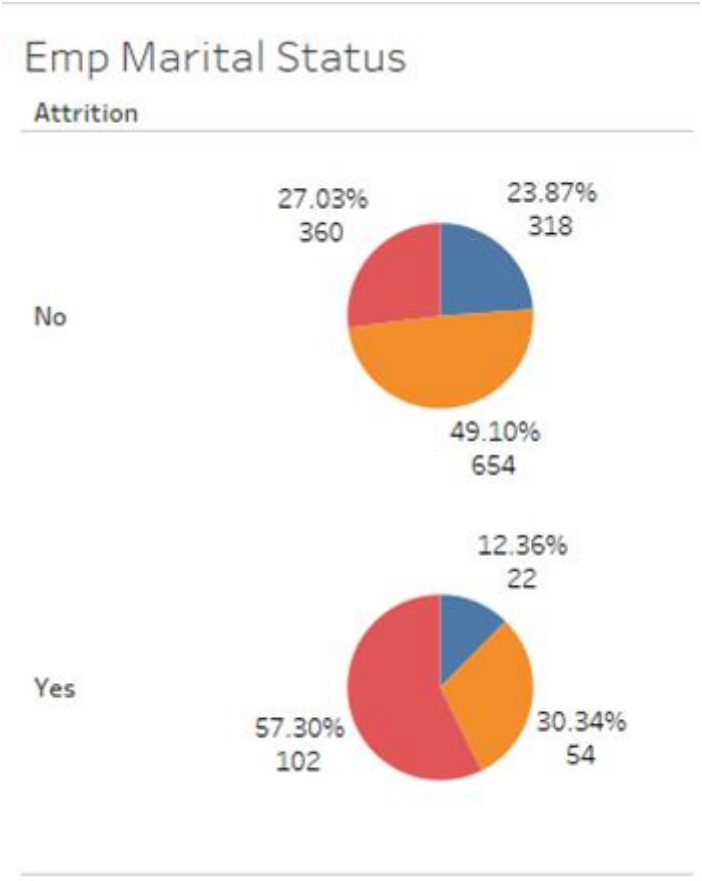
- Drag Overtime to Columns shelf.
- Drag Attrition to Rows shelf.
- Drag Count of Employee to Text shelf.
- Choose Quick Table Calculation > Percent of Total by right-clicking on the Count of Employees in Text shelf.
- Go to Show Me and select Pie Chart.

4.2.2 Age:



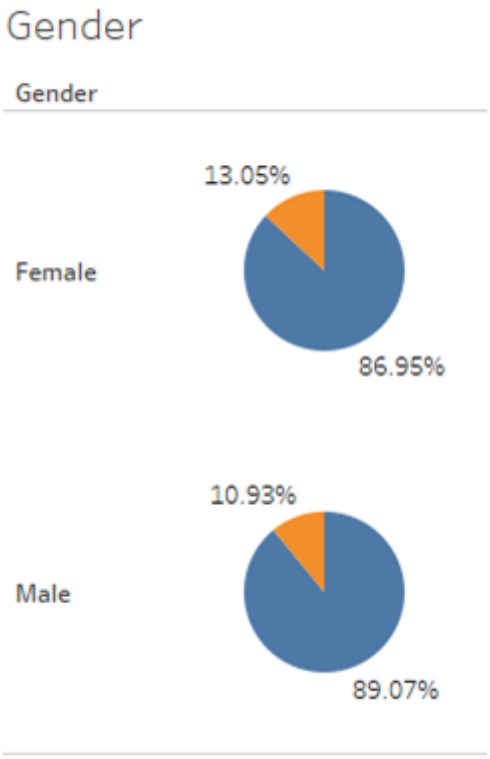
- Drag the Age variable over to the Rows shelf.
- To create bins based on age, right-click the age box, pick New > Bins, and then set the Bin Size to 5.
- Drag Attrition to Columns shelf.
- Drag Count of Employee to Text shelf.
- Drag Age (bin) to Columns shelf, next to Attrition.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Right-click on Count of Employee in Text shelf again and select Add to Label.

4.2.3 Marital Status:



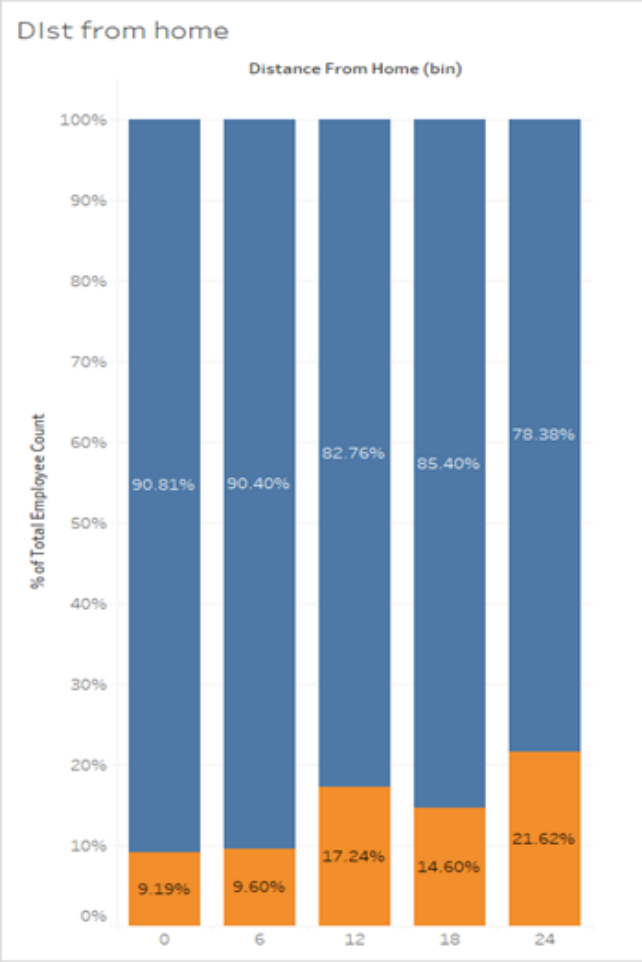
- Drag Marital Status to Columns shelf.
- Drag Attrition to Rows shelf.
- Drag Count of Employee to Text shelf.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Go to Show Me and select Pie Chart.

4.2.4 Gender:



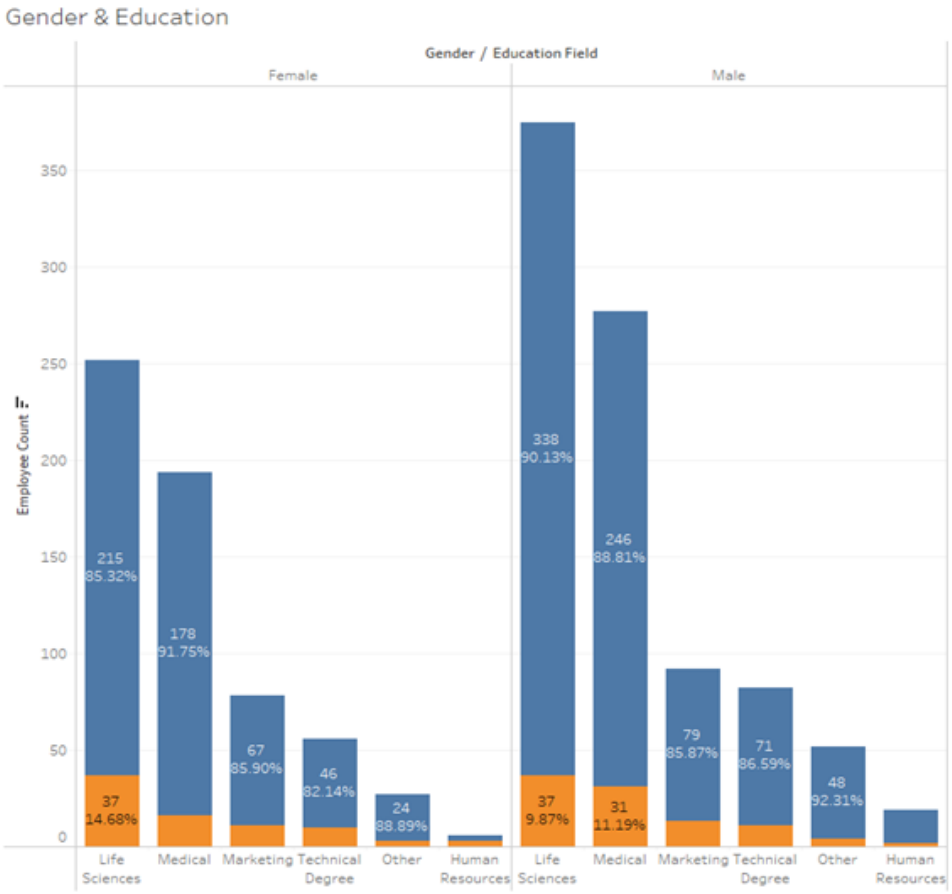
- Drag Gender to Columns shelf.
- Drag Attrition to Rows shelf.
- Drag Count of Employee to Text shelf.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Go to Show Me and select Pie Chart.

4.2.5 Dist from home:



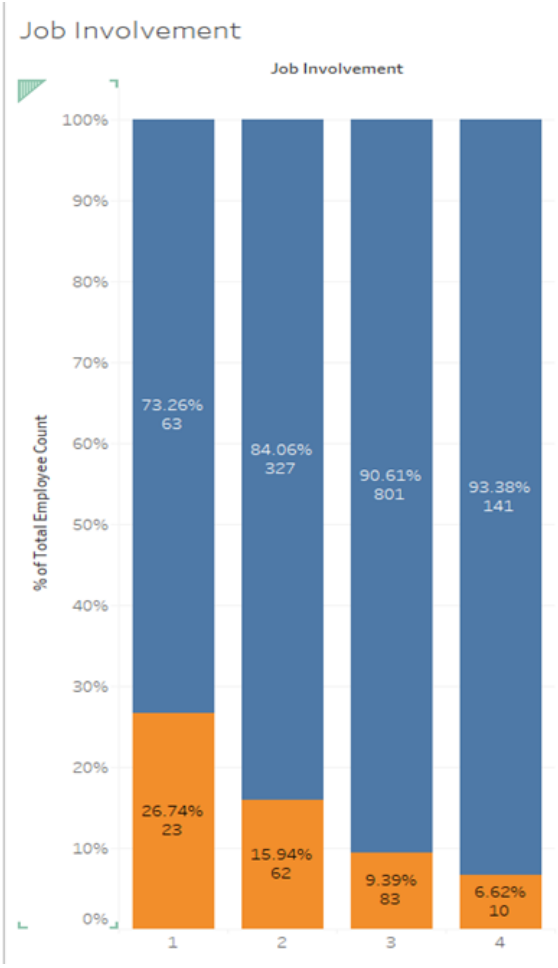
- Drag Distance from Home to Rows shelf.
- Right-click on Distance from Home field and select Create > Bins, set Size of bins to 6.
- Drag Attrition to Columns shelf.
- Drag Count of Employee to Text shelf.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Go to Show Me and select Circle View.

4.2.6 Gender & Ed:



- Drag Gender to Columns shelf.
- Drag Education to Columns shelf, next to Gender.
- Drag Count of Employee to Text shelf.
- Go to Show Me and select Stacked Column chart.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Right-click on Count of Employee in Text shelf again and select Add to Label.

4.2.7 Job Involvement:



- Drag Job Involvement to Columns shelf.
- Drag Attrition to Rows shelf.
- Drag Count of Employee to Text shelf.
- Go to Show Me and select Stacked Column chart.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Right-click on Count of Employee in Text shelf again and select Add to Label.

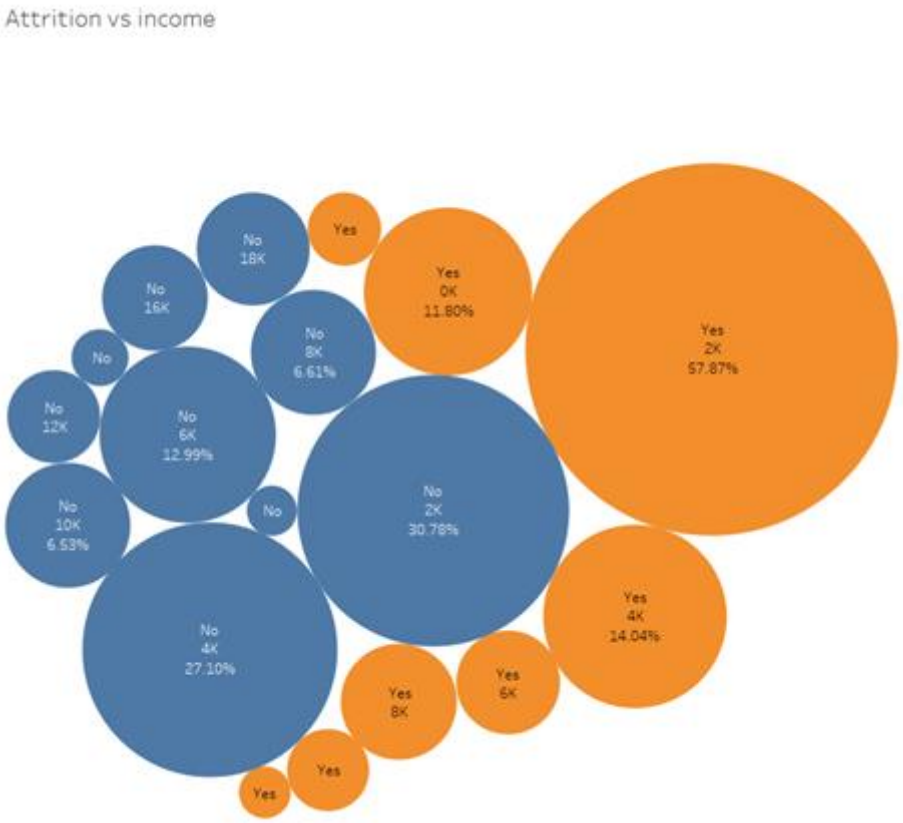
4.2.8 Years since last promotion:

Years since last promotion and Company Exp

Years At Co..	Attrition	Years Since Last Promotion (bin) 0	3	6	9	12	15
0	No	370					
	Yes	111					
4	No	382	36	20			
	Yes	35	4	3			
8	No	161	51	85	10		
	Yes	6	3	8	2		
12	No	19	33	14	9		
	Yes				3		
16	No	15	8	2	11	7	4
	Yes		1				
20	No	14	16	3	13	15	2
	Yes					1	
24	No	5	3		1	2	4
	Yes						
28	No	1			2	2	
	Yes					1	
32	No	4		1	2	1	2
	Yes						
36	No	1	1				
	Yes						

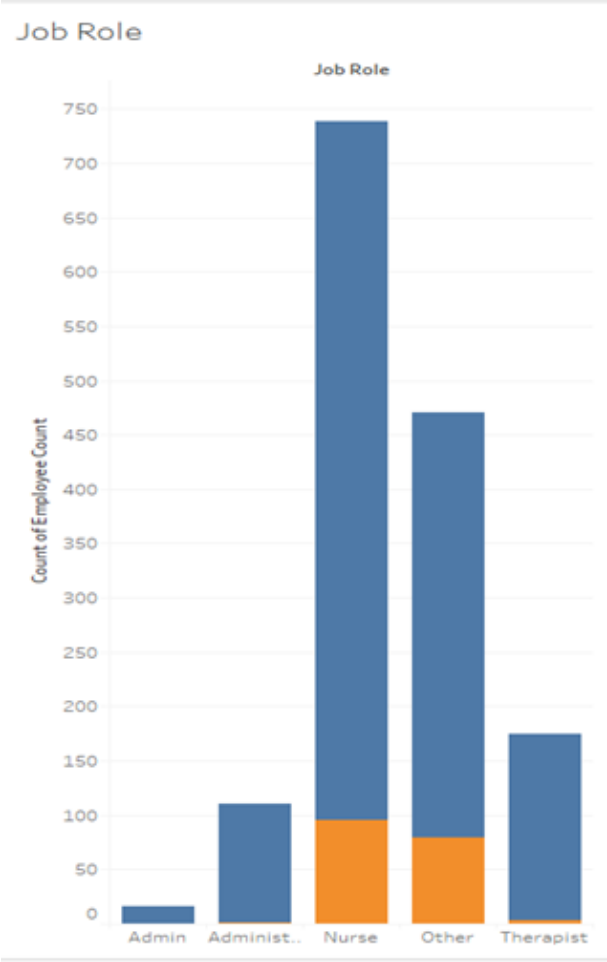
- Drag Years in current organization (bin) to Rows shelf.
- Drag Attrition to Columns shelf.
- Drag Years since last promotion (bin) to Columns shelf, next to Attrition.
- Drag Count of Employee to Text shelf.
- Select the highlighted table to create the plot.

4.2.9 Attrition vs income:



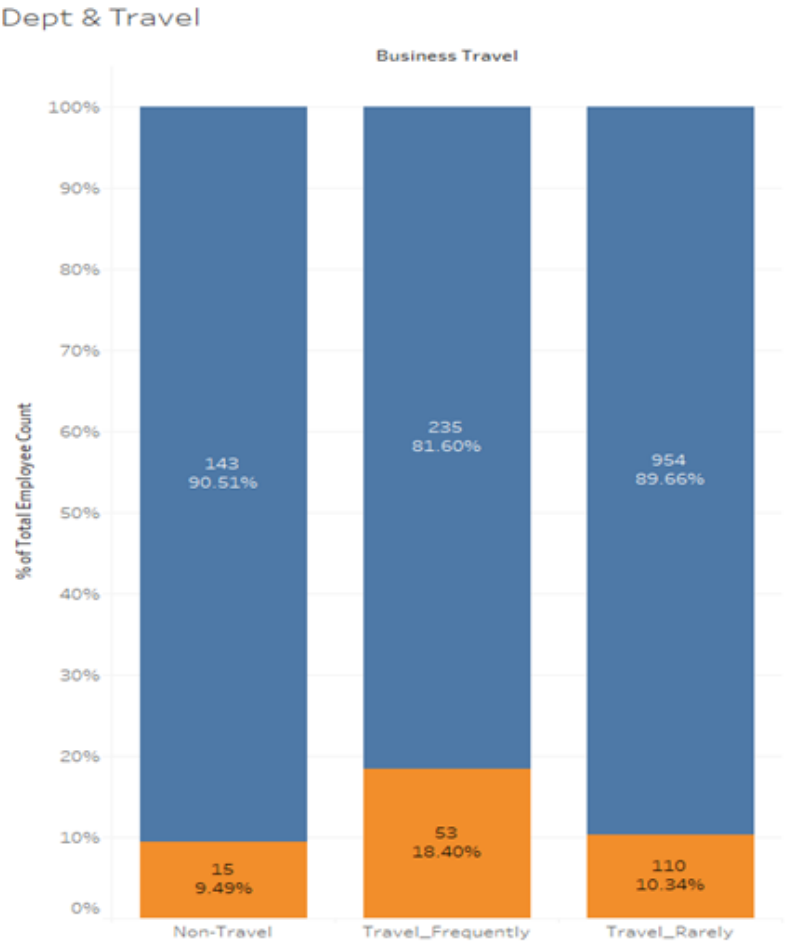
- Drag Income to Columns shelf.
- Right-click on Income field and select Create > Bins, set Size of bins to 2000.
- Drag Attrition to Rows shelf.
- Drag Count of Employee to Text shelf.
- Choose Quick Table Calculation > Percent of Total by right-clicking the Count of Employees in Text shelf.
- Drag Income (bin) to Size shelf.
- Go to Show Me and select Bubble Chart.

4.2.10 Job Role



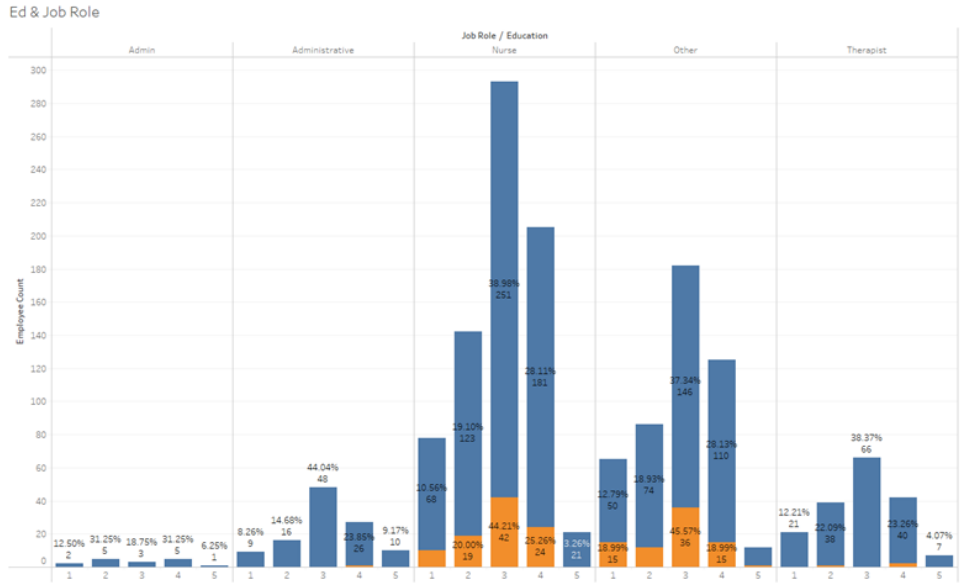
- Open your existing Tableau workbook and navigate to the worksheet where you want to create the Job Role visualization.
- Drag the Job Role field to the Columns shelf and the Attrition field to the Rows shelf.
- Drag the Emp count field to the Size or Color shelf.
- Change the chart type to Stacked Column chart from the Marks card.
- In the Label section, add the Emp count and the % of Emp count using the quick table calculation.

4.2.11 Department and Travel



- Open your existing Tableau workbook and navigate to the worksheet where you want to create the Department and Travel visualization.
- Drag the Department field to the Columns shelf and the Attrition variable into the Rows shelf.
- Drag the Emp count variable to the Size shelf.
- Change the chart type to Stacked Column chart from the Marks card.
- Choose Quick Table Calculation from the Analysis menu, and then choose Percent of Total.
- In label section, add the Emp count and the % of Emp count using the table calculation column wise.

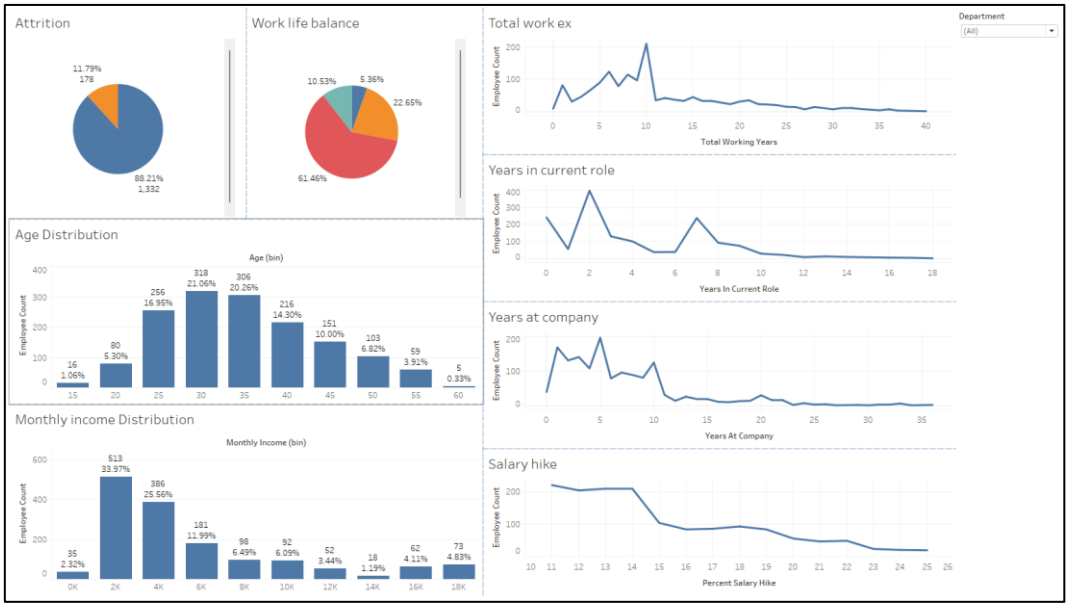
4.2.12 Education and Job Role

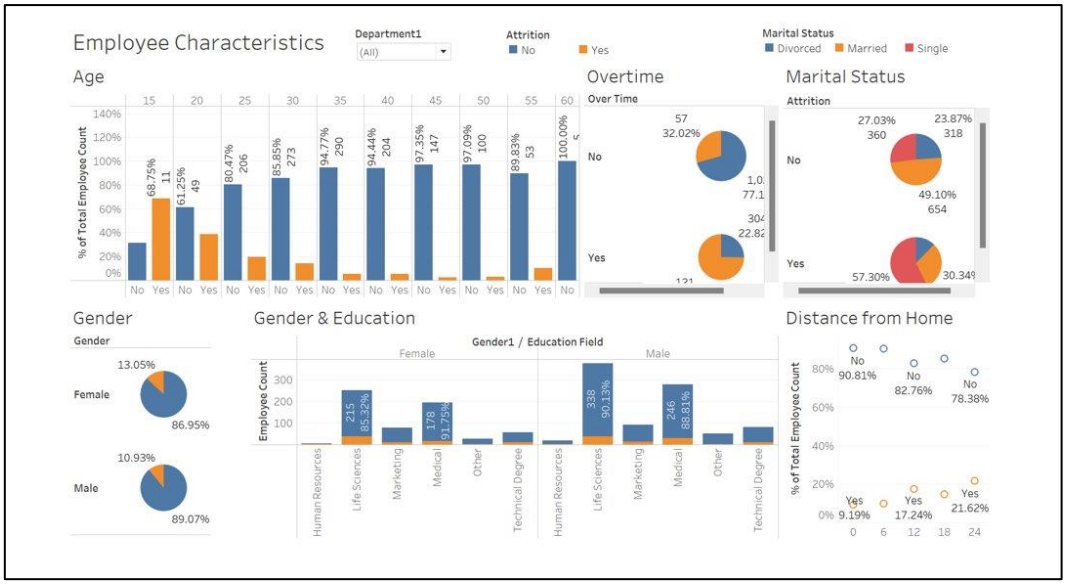


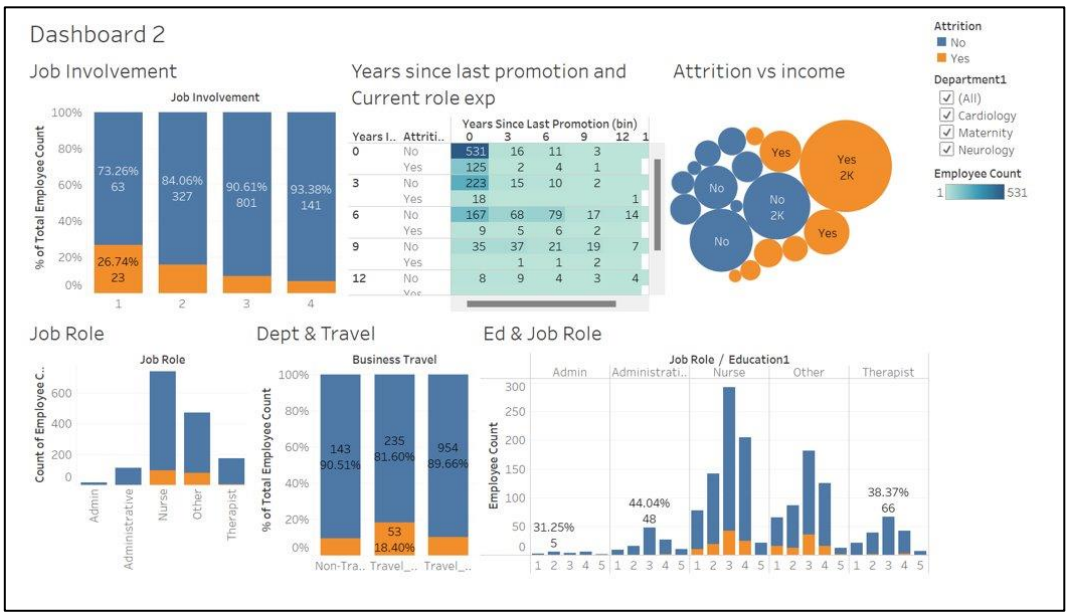
- Drag the Education Level field to the Columns shelf and the Job Role variable into the Rows shelf.
- Drag the Emp count variable to the Size shelf.
- Change the chart type to Stacked Column chart from the Marks card.
- Choose Quick Table Calculation from the Analysis menu, and then choose Percent of Total.
- In label section, add the Emp count and the % of Emp count using the table calculation column wise.

4.3 Dashboards

4.3.1 Organisation Level



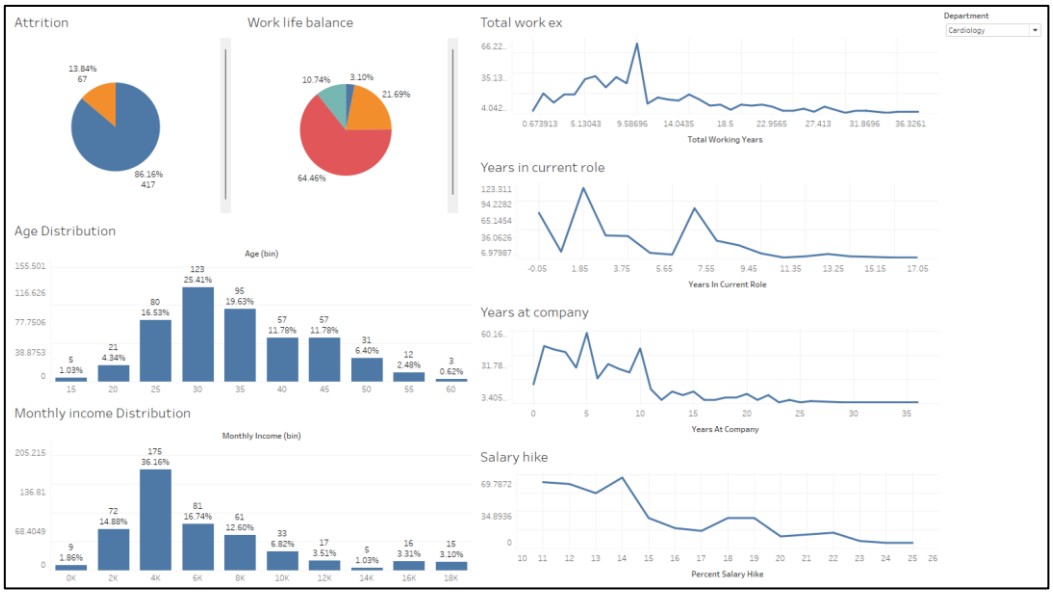


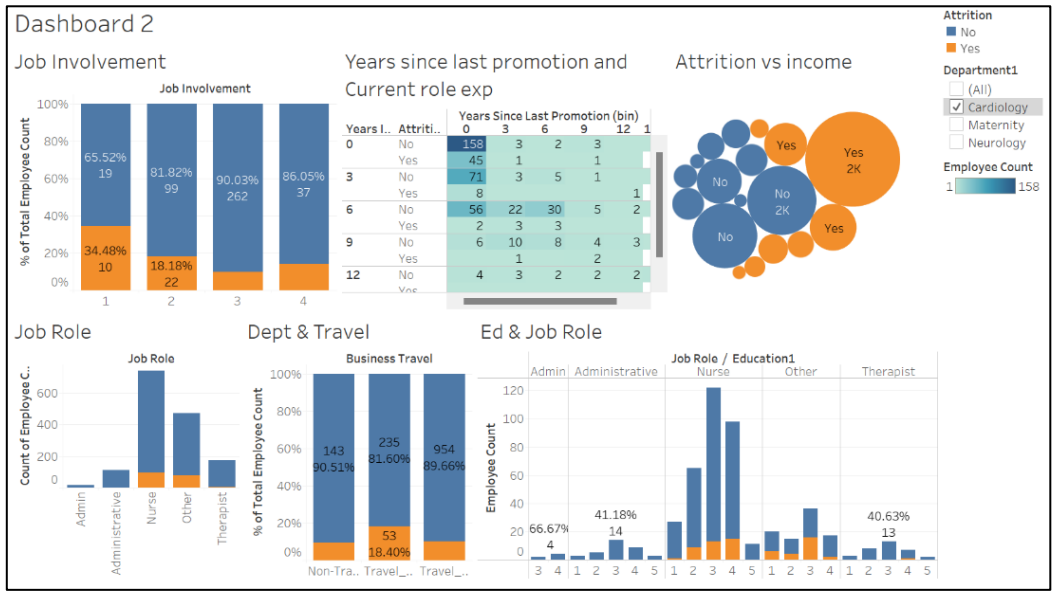


Based on the exploratory data analysis, the following insights can be drawn:

- The attrition rate for females is higher compared to males in the company.
- Single employees have a higher attrition rate compared to married employees, despite being in the minority.
- Employees who are required to do overtime have a higher probability of attrition.
- A greater distance from home is associated with a higher probability of an employee leaving the company.
- Younger employees are more likely to leave the company compared to those over the age of 40, who have a lower attrition rate.
- Employees with low job involvement are more likely to leave the company.
- Nurses and employees with other job roles have the highest attrition rate, whereas admin and administrative roles have an attrition rate of 0.
- Frequent travel increases the probability of attrition, compared to those who travel rarely or not at all.

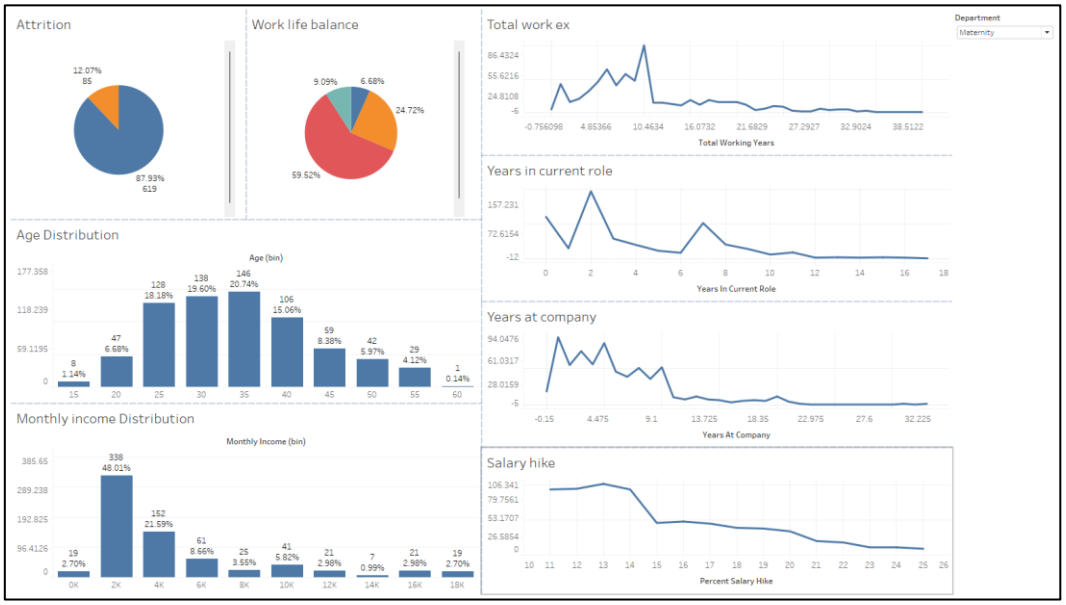
4.3.2 Department: Cardiology

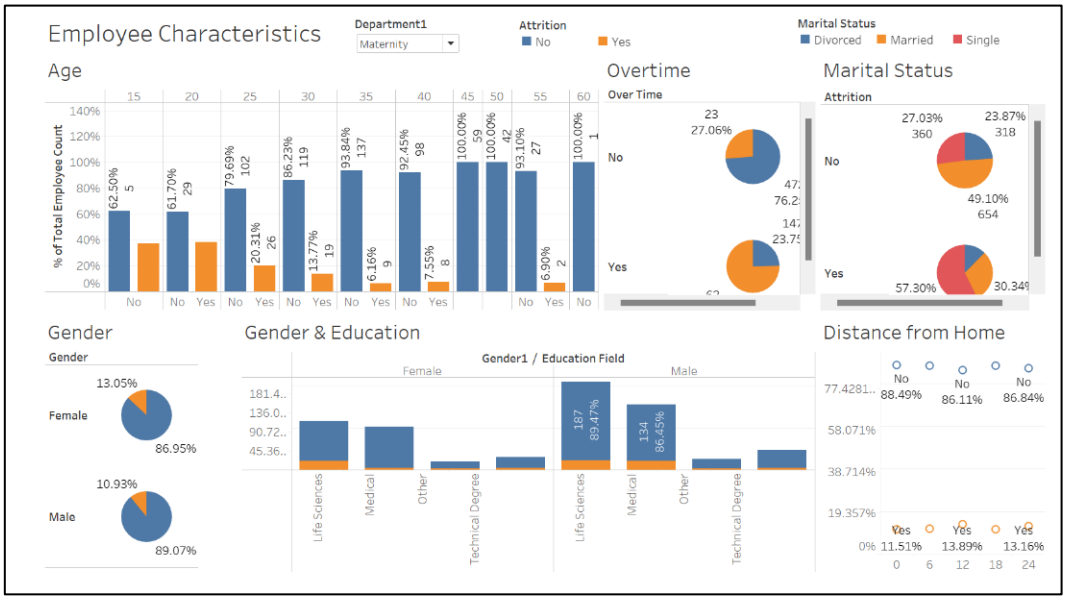


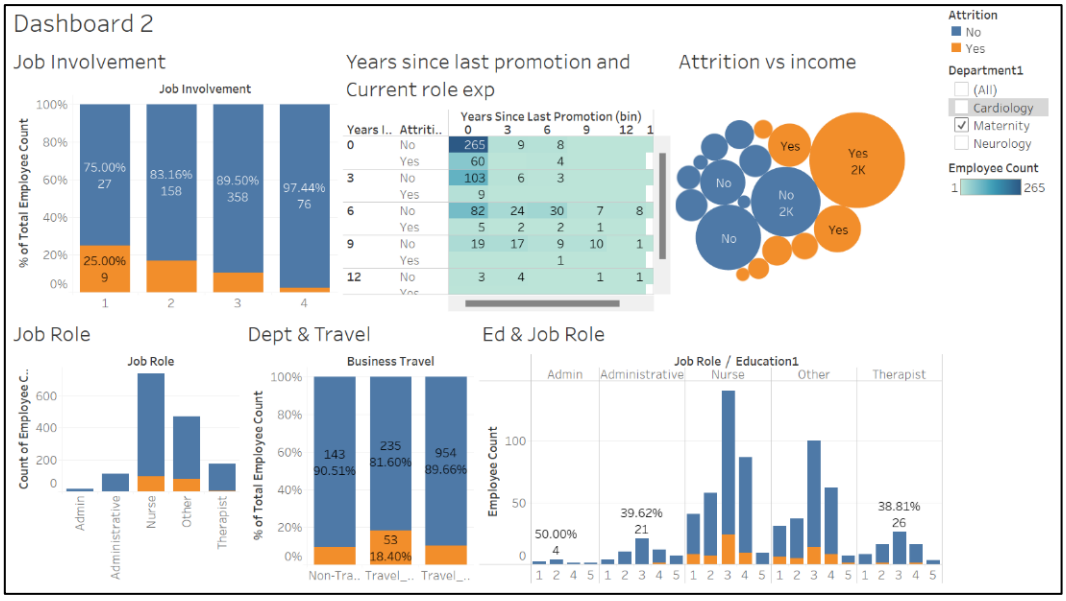


- Employees below the age of 20 have all left the job.
- Females have a higher attrition rate compared to males within their respective gender groups.
- An increase in the distance from home is associated with a significant increase in attrition rate.
- 57% of employees leaving the department have a salary lower than 2K.
- Most employees leaving are either nurses or from other departments, with an education level of 3 or 4.

4.3.3 Department: Maternity

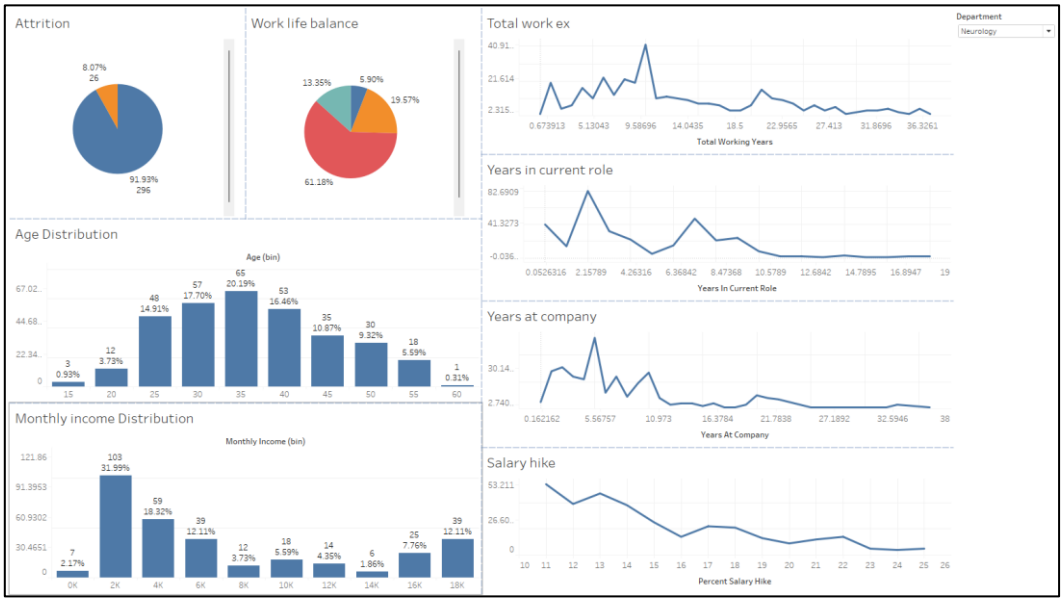


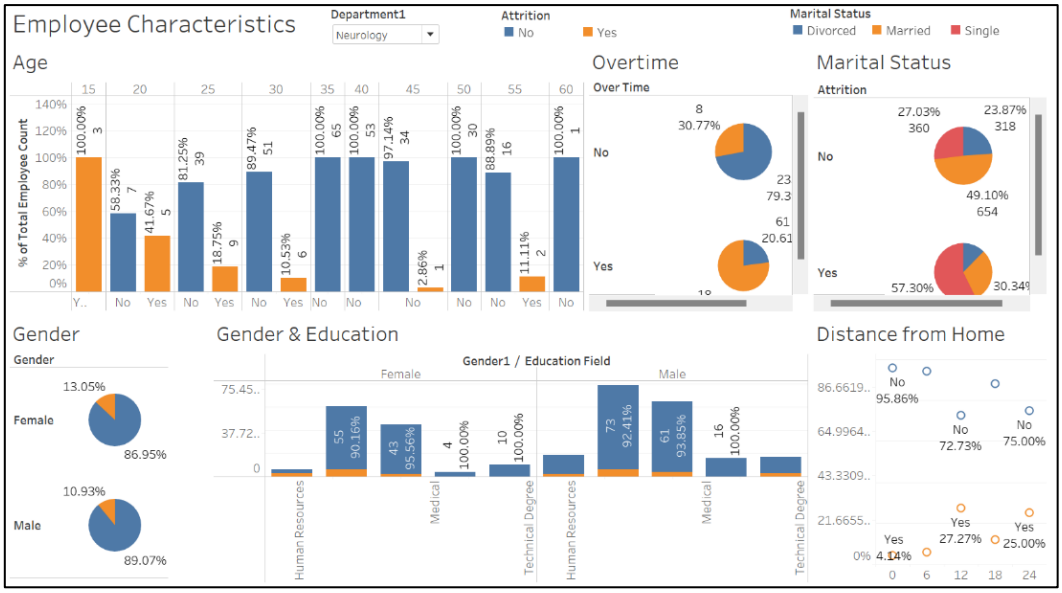


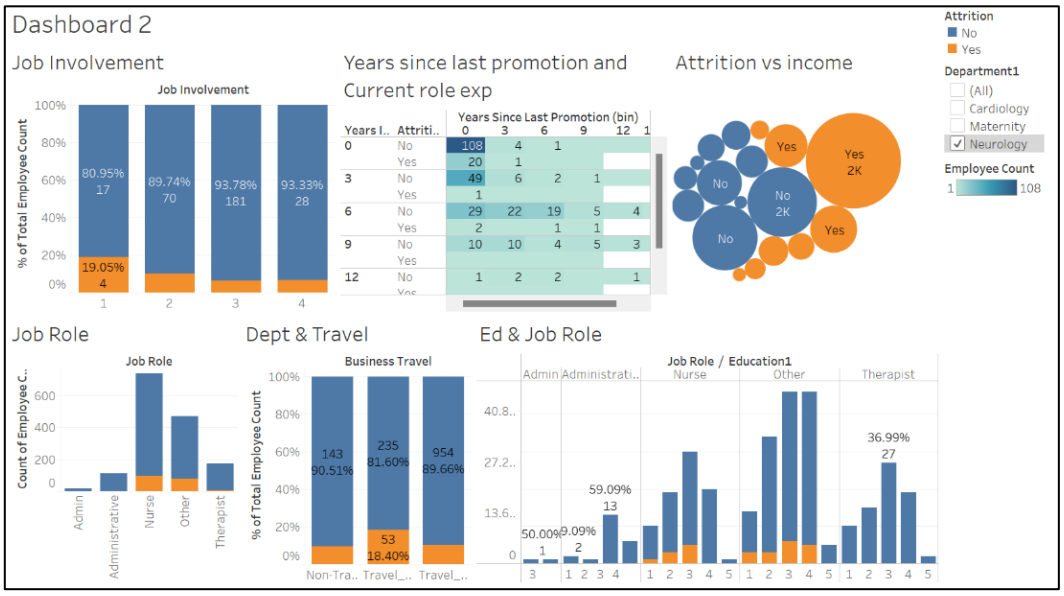


- Distance from home does not affect the attrition rate.
- Attrition rates are higher among employees with an educational background in life science and medical, regardless of gender.

4.3.4 Department: Neurology







- An increase in distance from home is associated with a significant increase in attrition rate.
- Employees over the age of 35 have a very low attrition rate.

5 Summary Statistics

Logistic Regression (Without Feature & Transformation)) › Summarize Data › Results dataset

rows 35 columns 23

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile	Mode
EmployeeID	1510	1510	0	1025177	1886378	1458584.027152	214399.408034	1238538.5	1469000.5	1668393.25	{1025177,1 1028630,1 ,1030937,' 1031524,1 032001,10 034220,10 1034448,1 1034814,1 1038656,1 041678,10 044017,10 044917,10 045879,1(
Age	1510	43	0	18	60	36.825828	7.360963	30	36	43	29
Attrition	1510	2	0								
BusinessTravel	1510	3	0								
DailyRate	1510	838	0	102	1499	793.242384	348.073409	459.5	787	1146	{329,444,'
Department	1510	3	0								
DistanceFromHome	1510	29	0	1	29	9.222517	6.634074	2	7	14	2
Education	1510	5	0	1	5	2.912583	0.791425	2	3	4	3
EducationField	1510	6	0								
EmployeeCount	1510	1	0	1	1	1	0	1	1	1	1

JobLevel	1510	5	0	1	5	2.072185	0.840939	1	2	3	1
JobRole	1510	5	0								
JobSatisfaction	1510	4	0	1	4	2.731126	0.959979	2	3	4	4
MaritalStatus	1510	3	0								
MonthlyIncome	1510	1227	0	1009	19999	6555.470199	3665.106066	2951.5	4907.5	8389.25	{2340,23
MonthlyRate	1510	1284	0	2097	26999	14268.405298	6248.009021	7817.25	14235.5	20461.5	{4223,19
NumCompaniesWorked	1510	10	0	0	9	2.675497	2.044463	1	2	4	1
Over18	1510	1	0								
OverTime	1510	2	0								
PercentSalaryHike	1510	15	0	11	25	15.157616	3.004112	12	14	18	11
PerformanceRating	1510	2	0	3	4	3.143709	0.246113	3	3	3	3
RelationshipSatisfaction	1510	4	0	1	4	2.697351	0.93404	2	3	4	3
StandardHours	1510	1	0	80	80	80	0	80	80	80	80
Shift	1510	4	0	0	3	0.817219	0.679753	0	1	1	0
TotalWorkingYears	1510	39	0	0	40	11.349007	6.099497	6	10	15	10
TrainingTimesLastYear	1510	7	0	0	6	2.817881	0.974444	2	3	3	2
WorkLifeBalance	1510	4	0	1	4	2.771523	0.539542	2	3	3	3
YearsAtCompany	1510	35	0	0	36	7.04702	4.443955	3	5	10	5
YearsInCurrentRole	1510	19	0	0	18	4.259603	3.035593	2	3	7	2
YearsSinceLastPromotion	1510	16	0	0	15	2.205298	2.363914	0	1	3	0
YearsWithCurrManager	1510	18	0	0	17	4.124503	3.020901	2	3	7	2

The descriptive analysis of the features is used to understand the nature of all the columns. There are no missing values in any of the columns of this dataset. Some of the important columns for our problem and their description:

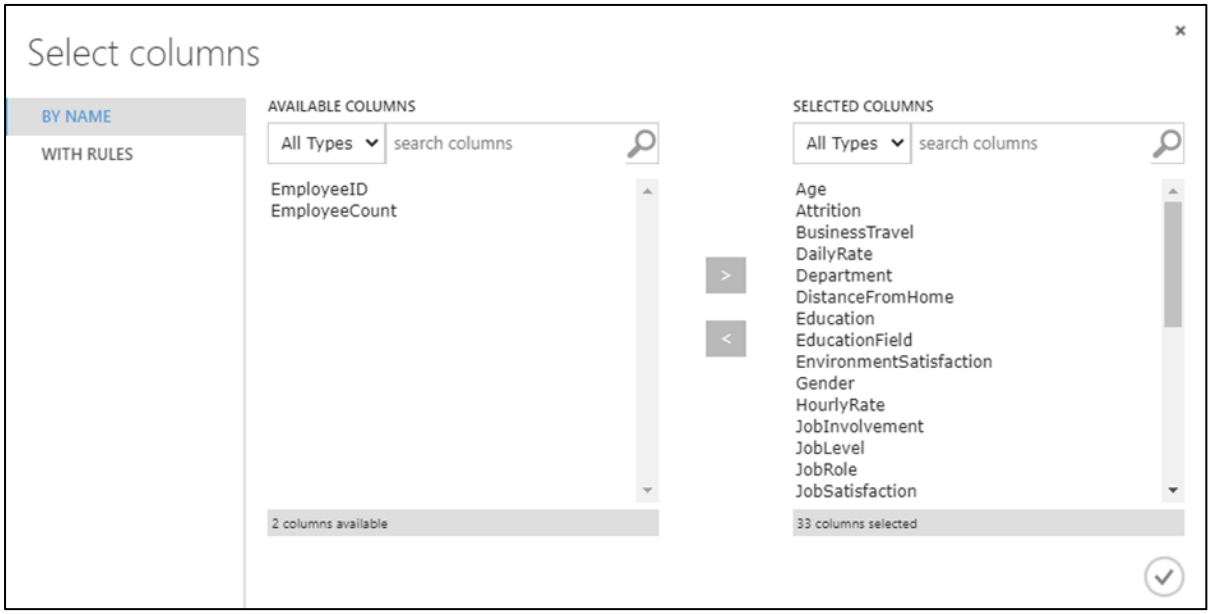
- Overtime – It describes whether an employee has worked overtime or not in terms of 'Yes' or 'No'
- TotalWorkingYears – It describes the career span of all the employees. Here the maximum no. of years for any employee goes up to 40 years while average working period is around 11 years. Majority employees have been working for 10 years.

- Age – describes the general age distribution of the employees which ranges between 18 to 60. The average age of the employees is around 36.8 years and majority employees are 29 years of age.
- MonthlyIncome – The monthly salary of the employee ranges from 1009 to 19999 depending on their job role. Average salary is around 3600 and majority employees are earning 2300 as monthly salary
- YearsAtCompany - It describes how long the employee has been working with the hospital which goes up to 36 years at max. Employees have been working for 7 years on average with the hospital and majority employees have been working for 5 years here.
- JobLevel – The job level is described in terms of 1-5, 1 being the lowest level. Average job level is 2 while most of the employees are working at the lowest level.
- YearsWithCurrManager – It describes the number of years an employee has been working under their current manager which goes up to a maximum and average working years is about 4.12 and majority employees have been working for 2 years with their current managers.

Here, Attrition is our dependent variable which is used to predict the attrition rate of the employees based on the above described features in the training dataset.

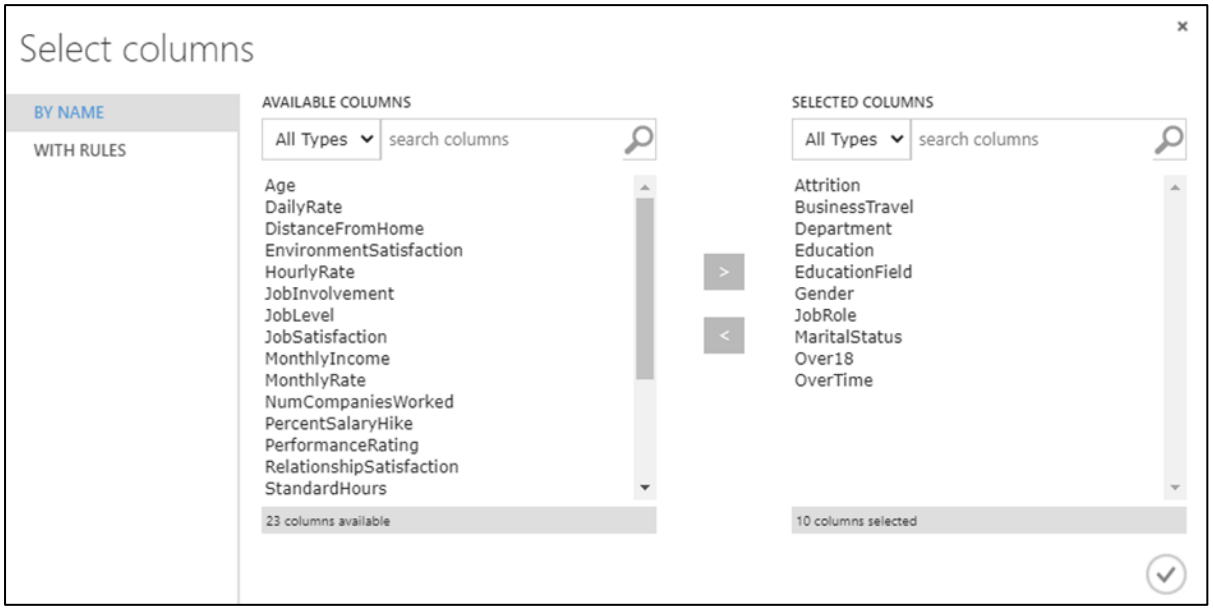
6 Data Pre-processing

6.1 Select Column



For the selection of columns, we considered 33 columns from available columns section apart from Employee ID & Employee Count

6.2 Edit Metadata

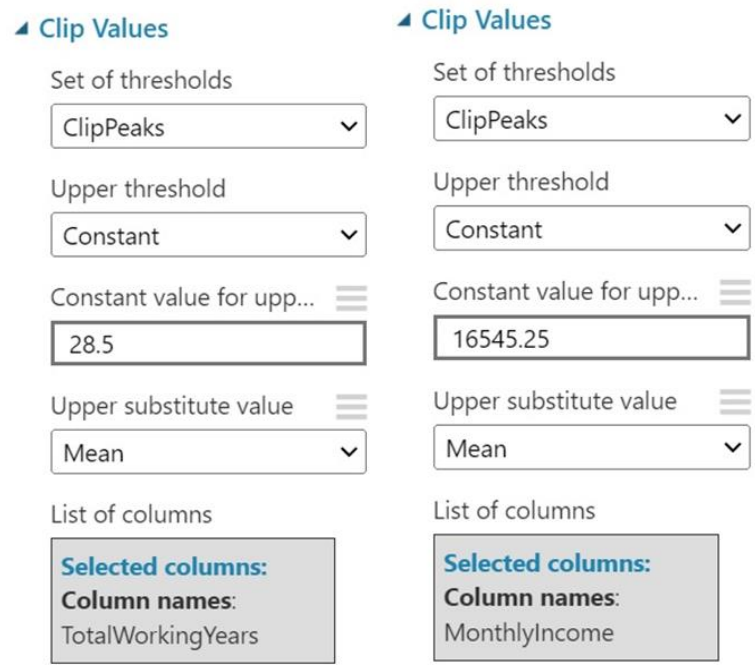


Edit metadata is used to change the definitions of Columns. Here the columns of string type were converted to categorical data type. 10 columns were added to metadata section.

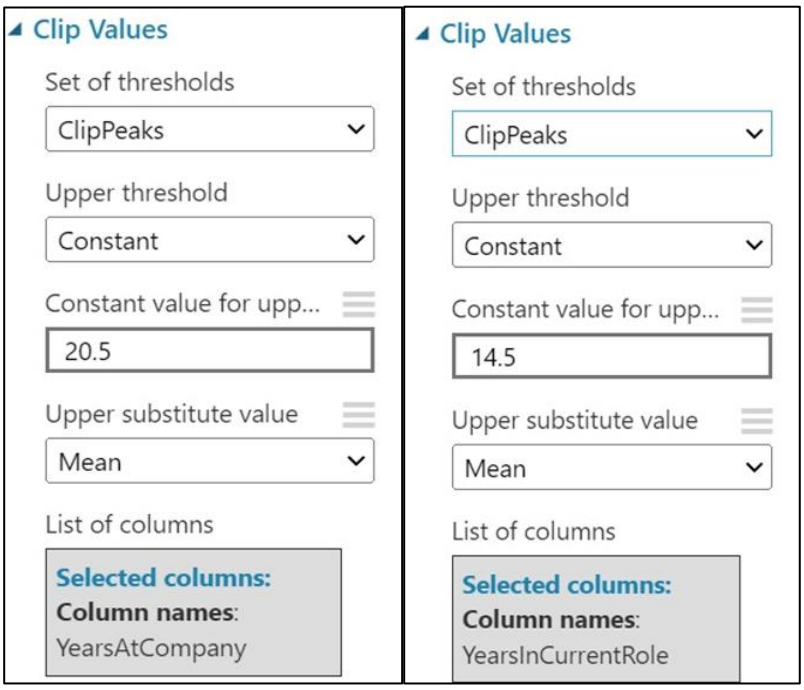
Feature	1st Quartile	3rd Quartile
MonthlyIncome	2951.5	8389.25
TotalWorkingYears	6	15
YearsAtCompany	3	10
YearsInCurrentRole	2	7

6.3 Outlier detection and treatment

Outliers were detected using IQR range which was multiplied by 1.5 and added it to the 3rd quartile. We found that features like Monthly income, years at company, years in current role and total working years were having outliers.



Clip value method was used to set the thresholds to treat the outliers. The outlier's greater than or equal to constant value was imputed by mean. The Constant value for upper threshold for the column 'Total Working Years' is 28.5 and constant value for upper threshold for the column 'Monthly Income' is 16545.25



The Constant value for upper threshold for the column 'Years at Company' is 20.5 and constant value for upper threshold for the column 'Years In current role' is 14.5.

6.4 Missing value treatment

As missing values lead to biased, or inaccurate results, missing value is a primary step in data pre -processing. Our summary statistical analysis revealed that there are no missing values in

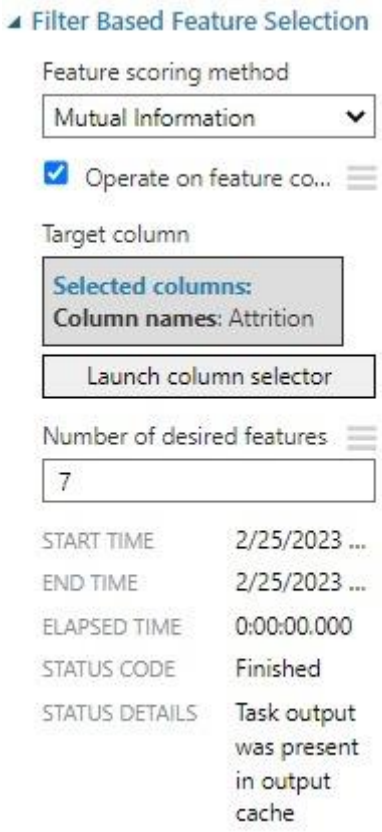
any of the columns in our dataset. So, the need to treat them or impute them was eliminated. The dataset was complete and ready for further analysis.

6.5 Correlation Matrix

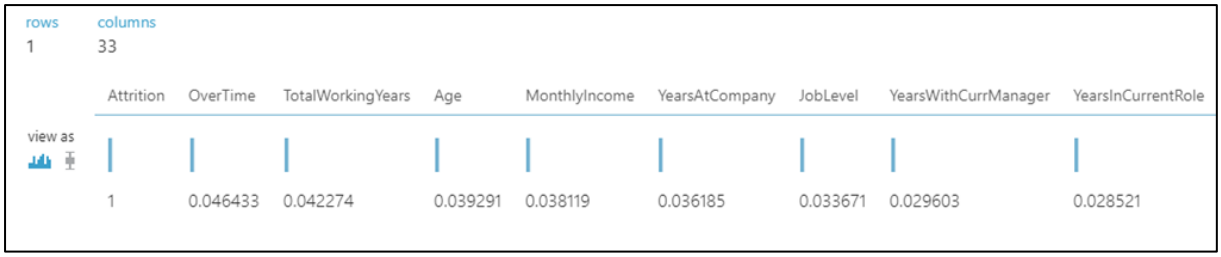
	Age	DailyRate	DistanceFr	Environme	HourlyRat	JobInvolv	JobLevel	JobSatisfa	MonthlyIn	MonthlyRa	NumComp	PercentSa	Performar	RelationsI	Shift	TotalWork	TrainingTi	WorkLifeB	YearsAtCo	YearsInCu	YearsSince	YearsWithCur
Age	1	-0.003119	-0.007914	0.02219	0.03604	0.037633	0.517317	-0.013291	0.406146	0.0309	0.287859	0.014789	-0.006306	0.062977	0.041344	0.575037	-0.011391	0.01218	0.222205	0.198039	0.220869	0.223254891
DailyRate	-0.003119	1	-0.005366	0.019825	0.038156	0.072519	0.008297	0.042854	0.011724	-0.040463	0.039839	0.013912	-0.010428	0.017654	0.06127	0.027089	0.006938	-0.020491	0.009345	0.012957	-0.039724	-0.020095198
DistanceFromHome	-0.007914	-0.005366	1	-0.025131	0.025382	0.001729	-0.019449	-0.013188	0.038891	0.037851	-0.015063	0.039018	0.0203	0.00058	0.032998	0.003113	-0.058382	-0.034789	0.028996	0.022186	-0.011162	-0.009596211
EnvironmentSatisfaction	0.02219	0.019825	-0.025131	1	-0.059749	-0.014332	0.018606	-0.008681	-0.009626	0.05923	0.029347	-0.01407	-0.020949	0.007942	-0.006844	-0.013686	-0.019777	0.02613	0.009059	0.04651	0.01042	0.004567866
HourlyRate	0.03604	0.038156	0.025382	-0.059749	1	0.029846	-0.007	-0.079976	0.011663	-0.021979	0.025512	-0.009882	0.001126	0.009105	0.041489	0.025376	-0.014395	-0.004973	-0.022896	-0.015495	-0.025809	-0.02060056
JobInvolvement	0.037633	0.072519	0.001729	-0.014332	0.029846	1	-0.008634	-0.038707	-0.01461	0.001982	0.026812	-0.012912	-0.016045	0.054124	0.024905	0.014079	-0.037036	-0.00089	0.037872	0.026262	-0.020019	0.047564219
JobLevel	0.517317	0.008297	-0.019449	0.018606	-0.007	-0.008634	1	-0.014865	0.726212	0.031834	0.158028	-0.007918	0.003953	0.026473	0.015875	0.68213	0.00131	0.055656	0.359605	0.319937	0.331939	0.372513158
JobSatisfaction	-0.013291	0.042854	-0.013188	-0.008681	-0.079976	-0.038707	-0.014865	1	-0.00377	-0.004867	-0.060436	0.049069	0.022782	-0.022745	0.023605	-0.020768	-0.002291	-0.016834	0.024635	0.005866	-0.013657	-0.025933088
MonthlyIncome	0.406146	0.011724	0.038891	-0.009626	0.011663	-0.01461	0.726212	-0.00377	1	0.05862	0.157887	0.032226	0.016816	0.011491	0.036631	0.553357	-0.032787	0.038516	0.363745	0.329258	0.266789	0.317955179
MonthlyRate	0.0309	-0.040463	0.037851	0.05923	-0.021979	0.001982	0.031834	-0.004867	0.05862	1	0.042282	-0.007309	-0.011912	-0.021033	-0.03733	-0.027897	0.004203	0.002368	-0.027687	-0.015267	0.000975	-0.049816551
NumCompaniesWorked	0.287859	0.039839	-0.015063	0.029347	0.025512	0.026812	0.158028	-0.060436	0.157887	0.042282	1	-0.009133	-0.024758	0.076215	0.019104	0.247179	-0.059279	0.000397	-0.096812	-0.093395	-0.022485	-0.091624321
PercentSalaryHike	0.014789	0.013912	0.039018	-0.01407	-0.009882	-0.012912	-0.007918	0.049069	0.032226	-0.007309	-0.009133	1	0.758761	-0.034749	0.012085	-0.003162	-0.029902	-0.004857	0.001553	0.026302	0.010227	-0.005155067
PerformanceRating	-0.006306	-0.010428	0.0203	-0.020949	0.001126	-0.016045	0.003953	0.022782	0.016816	-0.011912	-0.024758	0.758761	1	-0.033902	0.003664	0.021889	-0.035782	0.001556	0.037401	0.057713	0.046448	0.031958243
RelationshipSatisfaction	0.062977	0.017654	0.00058	0.007942	0.009105	0.054124	0.026473	-0.022745	0.011491	-0.021033	0.076215	-0.034749	-0.033902	1	-0.033405	0.017317	0.00275	0.015395	-0.011319	-0.032204	0.03465	0.001892727
Shift	0.041344	0.06127	0.032998	-0.006844	0.041489	0.024905	0.015875	0.023605	0.036631	-0.03733	0.019104	0.012085	0.003664	-0.033405	1	0.03173	0.026759	0.003232	0.021808	0.047319	0.009728	0.009890631
TotalWorkingYears	0.575037	0.027089	0.003113	-0.013686	0.025376	0.014079	0.68213	-0.020768	0.553357	-0.027897	0.247179	-0.003162	0.021889	0.017317	0.03173	1	0.008828	0.032225	0.500896	0.415408	0.378333	0.451899801
TrainingTimesLastYear	-0.011391	0.006938	-0.058382	-0.019777	-0.014395	-0.037036	0.00131	-0.002291	-0.032787	0.004203	-0.059279	-0.029902	-0.035782	0.00275	0.026759	0.008828	1	0.019817	-0.00664	0.004393	-0.005008	-0.001846693
WorkLifeBalance	0.01218	-0.020491	-0.034789	0.02613	-0.004973	-0.00089	0.055656	-0.016834	0.038516	0.002368	0.000397	-0.004857	0.001556	0.015395	0.003232	0.032225	0.019817	1	0.028373	0.047825	0.026206	0.011952021
YearsAtCompany	0.222205	0.009345	0.028996	0.009059	-0.022896	0.037872	0.359605	0.024635	0.363745	-0.027687	-0.096812	0.001553	0.037401	-0.011319	0.021808	0.500896	-0.00664	0.028373	1	0.769344	0.51811	0.782713554
YearsInCurrentRole	0.198039	0.012957	0.022186	0.04651	-0.015495	0.026262	0.319937	0.005866	0.329258	-0.015267	-0.093395	0.026302	0.057713	-0.032204	0.047319	0.415408	0.004393	0.047825	0.769344	1	0.509834	0.695460212
YearsSinceLastPromotion	0.220869	-0.039724	-0.011162	0.01042	-0.025809	-0.020019	0.331939	-0.013657	0.266789	0.000975	-0.022485	0.010227	0.046448	0.03465	0.009728	0.378333	-0.005008	0.026206	0.51811	0.509834	1	0.520801678
YearsWithCurrManager	0.223255	-0.020095	-0.009596	0.004568	-0.020601	0.047564	0.372513	-0.025933	0.317955	-0.049817	-0.091624	-0.005155	0.031958	0.001893	0.009891	0.4519	-0.001847	0.011952	0.782714	0.69546	0.520802	1

Fig: Used correlation statistical function to compute the correlation matrix. Red colour indicates high correlation and green colour indicates inverse correlation.

6.6 Feature Selection



The scoring method used here is Mutual Information. Mutual information is used because it measures the degree of association between feature and target variable based on the information the feature provides about the target.

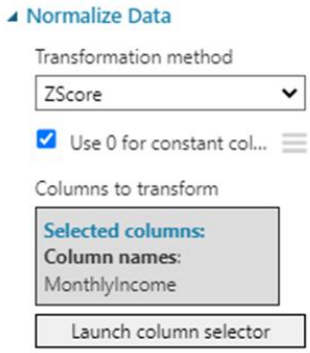


rows	columns
1	33

	Attrition	OverTime	TotalWorkingYears	Age	MonthlyIncome	YearsAtCompany	JobLevel	YearsWithCurrManager	YearsInCurrentRole
view as									
	1	0.046433	0.042274	0.039291	0.038119	0.036185	0.033671	0.029603	0.028521

The top 7 features were selected based on the ranking of the features. The features include Over Time, Total Working Hours, Age, Monthly Income, Years at Company, Job Level, Years with Current Manager.

6.7 Normalisation using Z score

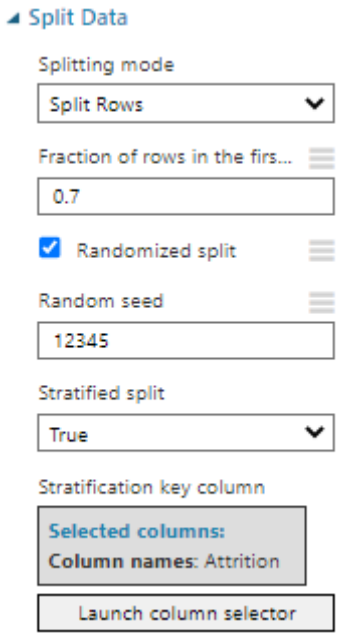


Normalization to a z-score was used as the transformation for this dataset. The goal of the statistical procedure known as Z-score normalisation (or standardisation) is to modify a data set such that its new mean is zero and its new standard deviation is one. To do this, first remove the data's mean from each individual point, and then divide that total by the standard deviation.

The final converted data set has a mean of 0 and a standard deviation of 1, making it simpler to compare data points that originally had various units or scales. It's helpful when dealing with data that has varying units of measurement, such as monthly salary, age, job level, overtime, etc.

7 Train Model

7.1 Split Data

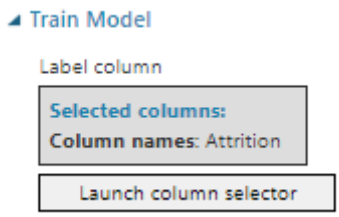


We have first split the entire dataset in 9:1 ratio where 10% of the data has been kept aside for validation and the remaining 90% of the data further split into 7: 3 ratios for train-test process.

We have set the random seed to 12345 so that the same result will be reproduced. As machine learning model is trained, it involves several random processes, such as initializing weights, shuffling data, and selecting samples for validation. If these random processes are not controlled, the same model may produce different results each time it is trained on the same dataset. Hence, its necessary to set the random seed.

Also, the stratified split is set to TRUE to avoid class imbalance. When splitting a dataset into training and testing subsets, it is important to ensure that the distribution of the target variable is maintained in both subsets. This is especially important in classification problems where the classes may be imbalanced, i.e., some classes have significantly fewer samples than others. In such cases, a simple random split may lead to one or more classes being underrepresented in either the training or testing subset, which can negatively impact the model's performance. Stratified split ensures that the distribution of the target variable is maintained in both subsets by randomly selecting samples from each class such that the proportions of the classes are preserved in both subsets. This helps to ensure that the model is trained on a representative sample of the data and that its performance is evaluated on a testing dataset that is representative of the entire population.

7.2 Train Model



In train model, we set the dependent variable to Attrition.

8 Predictive Analytics using Azure

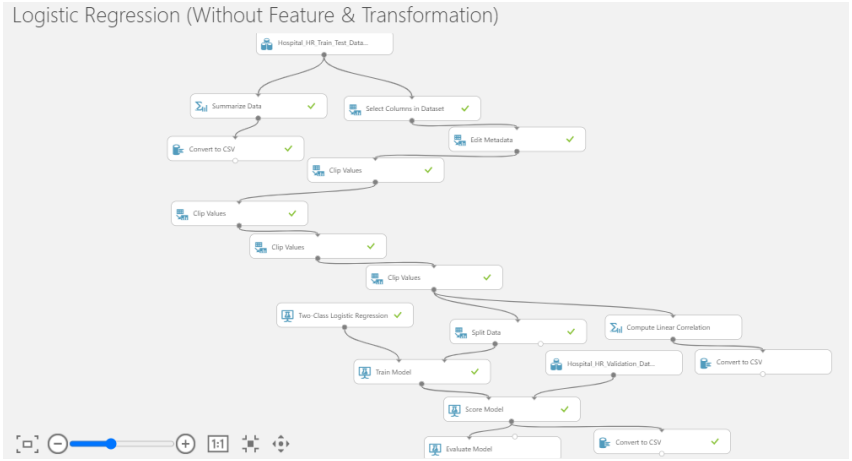
8.1 Model 1: Logistic Regression

A statistical technique known as logistic regression is used to examine the association between a categorical dependent variable and a set of independent factors. The dependent variable (Attrition) is a binary indicator of whether or not an employee has left the firm, and the independent variables are numerous characteristics that may impact an employee's choice to leave, such as work satisfaction, compensation, and tenure.

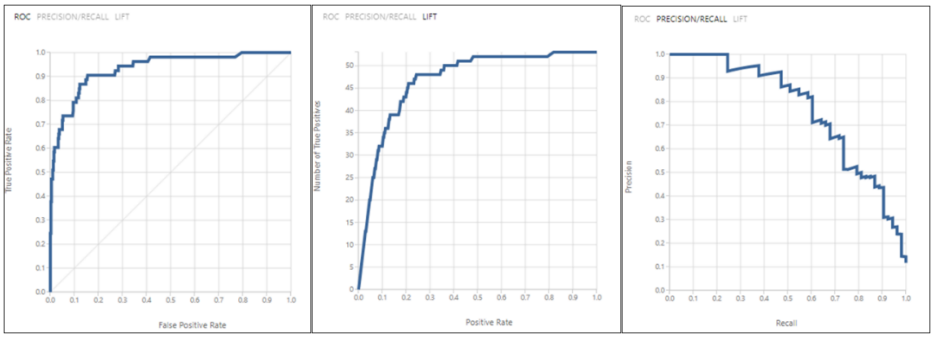
Once the data is collected, the next step would be to pre-process the data, which includes cleaning the data, handling missing values, feature selection and transforming the data into a format suitable for analysis.

A binary logistic regression approach, which evaluates the connection between the explanatory factors and the response, would then be used to train the logistic regression model using the cleaned and prepared data. As a function of the input variables, the model calculates the likelihood of an employee quitting the company.

8.1.1 Without Feature selection & Transformation



The model can be evaluated using various performance metrics, such as accuracy, precision, recall, and F1 score.



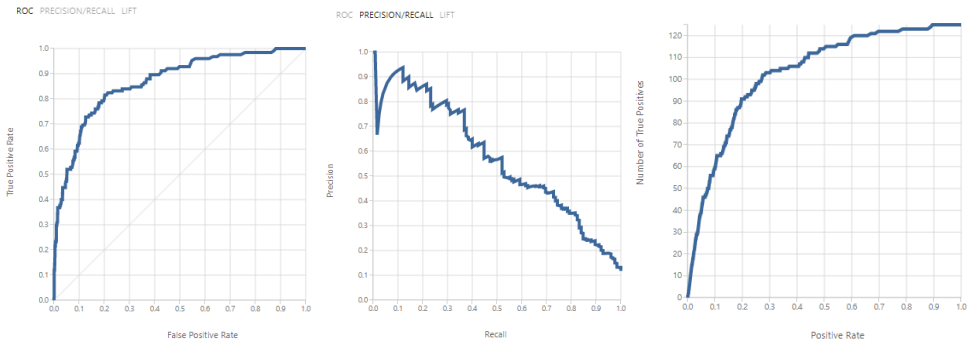
Evaluation of non-feature and without transformation logistic regression model:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
27	26	0.932	0.844	0.5	0.933
False Positive	True Negative	Recall	F1 Score		
5	395	0.509	0.635		
Positive Label	Negative Label				
Yes	No				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	4	0	0.009	0.892	0.140	1.000	0.075	0.891	1.000	0.000
(0.800,0.900]	5	0	0.020	0.903	0.290	1.000	0.170	0.901	1.000	0.000
(0.700,0.800]	5	1	0.033	0.912	0.412	0.933	0.264	0.911	0.998	0.001
(0.600,0.700]	6	1	0.049	0.923	0.533	0.909	0.377	0.923	0.995	0.002
(0.500,0.600]	7	3	0.071	0.932	0.635	0.844	0.509	0.938	0.988	0.005
(0.400,0.500]	5	5	0.093	0.932	0.674	0.762	0.604	0.949	0.975	0.013
(0.300,0.400]	4	9	0.121	0.921	0.667	0.655	0.679	0.957	0.953	0.027
(0.200,0.300]	6	20	0.179	0.890	0.627	0.519	0.792	0.970	0.902	0.064
(0.100,0.200]	6	36	0.272	0.823	0.545	0.390	0.906	0.985	0.813	0.142
(0.000,0.100]	5	325	1.000	0.117	0.209	0.117	1.000	1.000	0.000	0.933

8.1.2 With Feature selection only

Evaluation of feature selected logistic regression model:

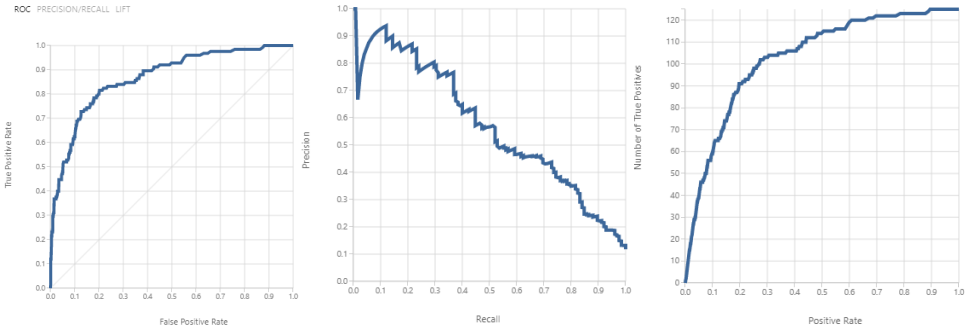


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
14	39	0.903	0.737	0.5	0.846
False Positive	True Negative	Recall	F1 Score		
5	395	0.264	0.389		
Positive Label	Negative Label				
Yes	No				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.883	0.000	1.000	0.000	0.883	1.000	0.000
(0.800,0.900]	0	0	0.000	0.883	0.000	1.000	0.000	0.883	1.000	0.000
(0.700,0.800]	4	0	0.009	0.892	0.140	1.000	0.075	0.891	1.000	0.000
(0.600,0.700]	4	1	0.020	0.898	0.258	0.889	0.151	0.899	0.998	0.000
(0.500,0.600]	6	4	0.042	0.903	0.389	0.737	0.264	0.910	0.988	0.002
(0.400,0.500]	5	9	0.073	0.894	0.442	0.576	0.358	0.919	0.965	0.009
(0.300,0.400]	3	10	0.102	0.879	0.444	0.478	0.415	0.924	0.940	0.019
(0.200,0.300]	9	19	0.163	0.857	0.488	0.419	0.585	0.942	0.892	0.042
(0.100,0.200]	11	61	0.322	0.746	0.422	0.288	0.792	0.964	0.740	0.152
(0.000,0.100]	11	296	1.000	0.117	0.209	0.117	1.000	1.000	0.000	0.846

Evaluation of Logistic Regression using feature selection and transformation applied:

8.1.3 With both Feature selection & Transformation



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
14	39	0.903	0.737	0.5	0.846
False Positive	**True Negative**	**Recall**	**F1 Score**		
5	395	0.264	0.389		
Positive Label	**Negative Label**				
Yes	No				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.883	0.000	1.000	0.000	0.883	1.000	0.000
(0.800,0.900]	0	0	0.000	0.883	0.000	1.000	0.000	0.883	1.000	0.000
(0.700,0.800]	4	0	0.009	0.892	0.140	1.000	0.075	0.891	1.000	0.000
(0.600,0.700]	4	1	0.020	0.898	0.258	0.889	0.151	0.899	0.998	0.000
(0.500,0.600]	6	4	0.042	0.903	0.389	0.737	0.264	0.910	0.988	0.002
(0.400,0.500]	5	9	0.073	0.894	0.442	0.576	0.358	0.919	0.965	0.009
(0.300,0.400]	3	10	0.102	0.879	0.444	0.478	0.415	0.924	0.940	0.019
(0.200,0.300]	9	19	0.163	0.857	0.488	0.419	0.585	0.942	0.892	0.042
(0.100,0.200]	11	61	0.322	0.746	0.422	0.288	0.792	0.964	0.740	0.152
(0.000,0.100]	11	296	1.000	0.117	0.209	0.117	1.000	1.000	0.000	0.846

8.2 Model 2: Decision Tree

In machine learning, a decision tree is a common technique for making predictions by repeatedly segmenting the data into subsets according to the values of various attributes. It is possible to utilise a decision tree to determine which criteria, like as income, job satisfaction, and length of service, have the most impact on employee turnover rates in the context of attrition forecasting.

To build a decision tree to predict attrition rates, the first step would be to collect data on the independent variables (factors influencing attrition) and the dependent variable (whether an employee has left the organization or not) for a sample of employees. The data would then be split into a training set and a test set.

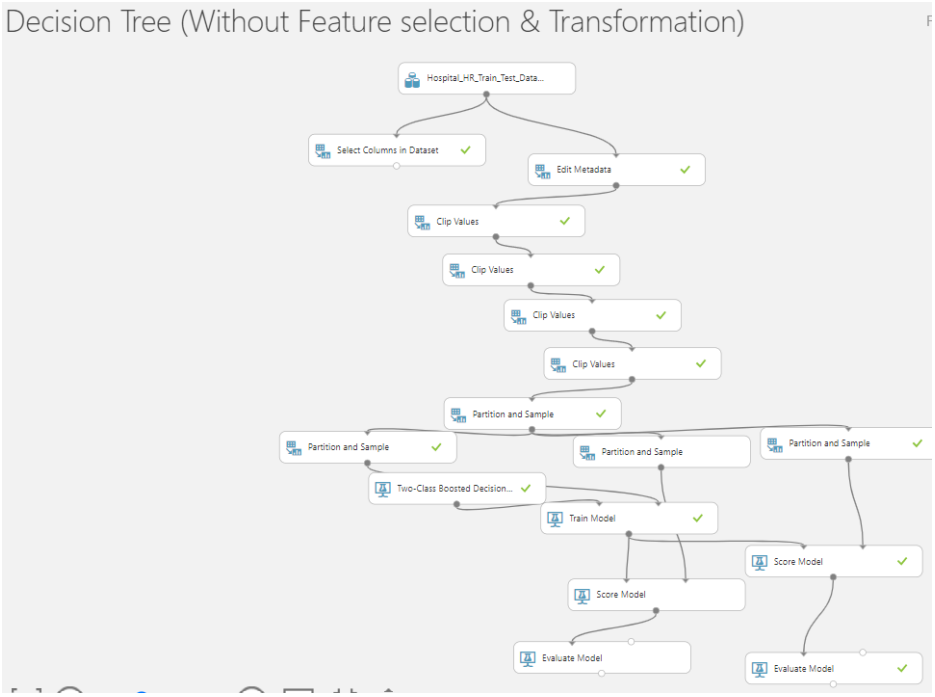
The decision tree algorithm would then be applied to the training set, the method builds a tree-like model that repeatedly divides the data into subsets according to the values of the independent variables. At each split, the algorithm selects the feature that best separates the data into the two classes (employees who left and employees who didn't) based on an impurity measure, such as the Gini impurity or entropy.

Once the decision tree is built, it can be used to predict the attrition rate of new employees based on their values of the independent variables. For example, if the decision tree splits on job satisfaction, salary, and tenure, a new employee's values for these features can be used to traverse the tree and predict whether they are likely to leave the organization or not.

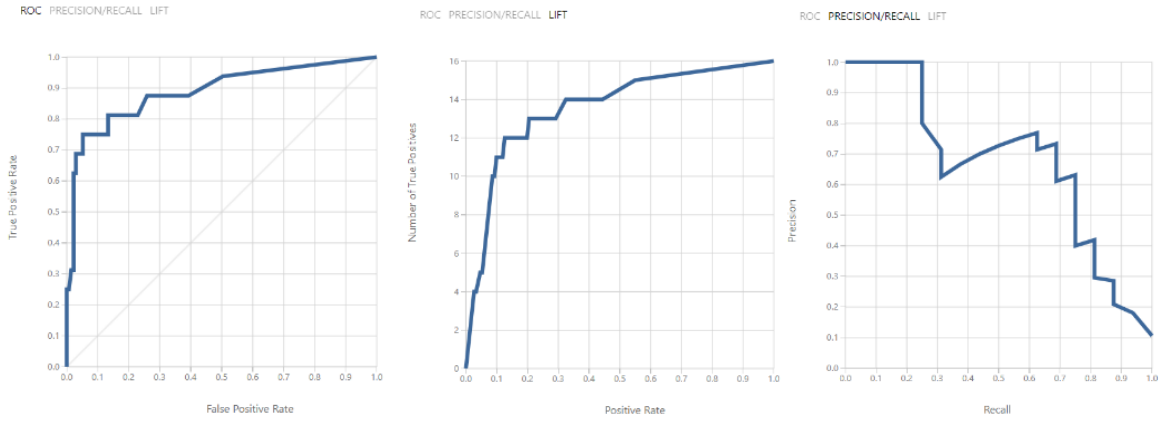
Many measures, including accuracy, precision, recall, and F1 score, may be used to determine the quality of the decision tree model's predictions. The model can also be tuned

by adjusting hyperparameters, such as the maximum depth of the tree, to improve its performance.

8.2.1 Without Feature selection & Transformation



Performance Metrics:



Evaluation:

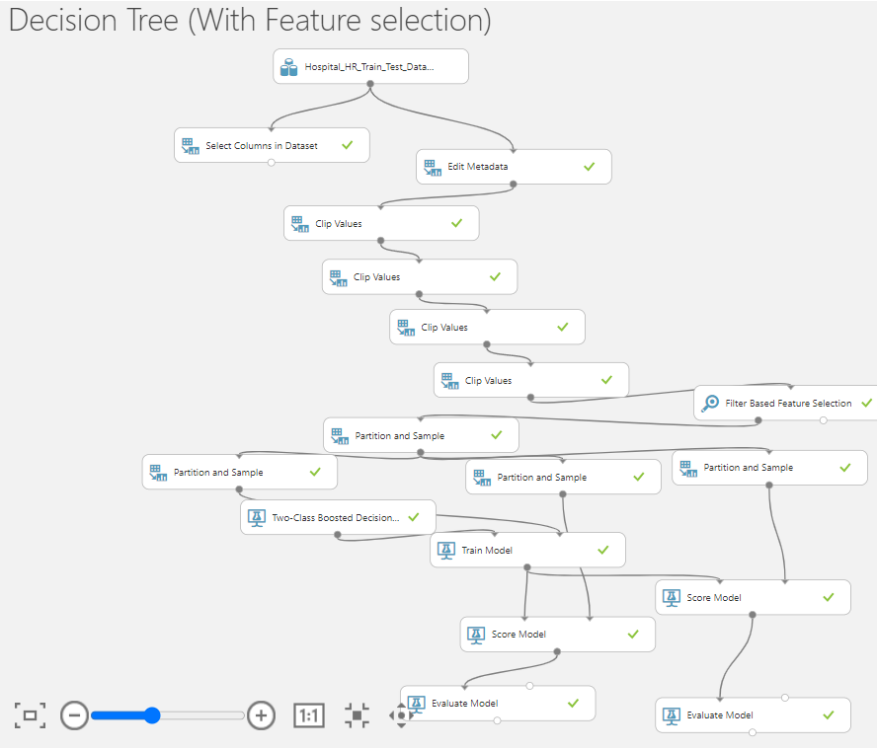
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
10	6	0.940	0.769	0.5	0.875

False Positive	True Negative	Recall	F1 Score
3	132	0.625	0.690

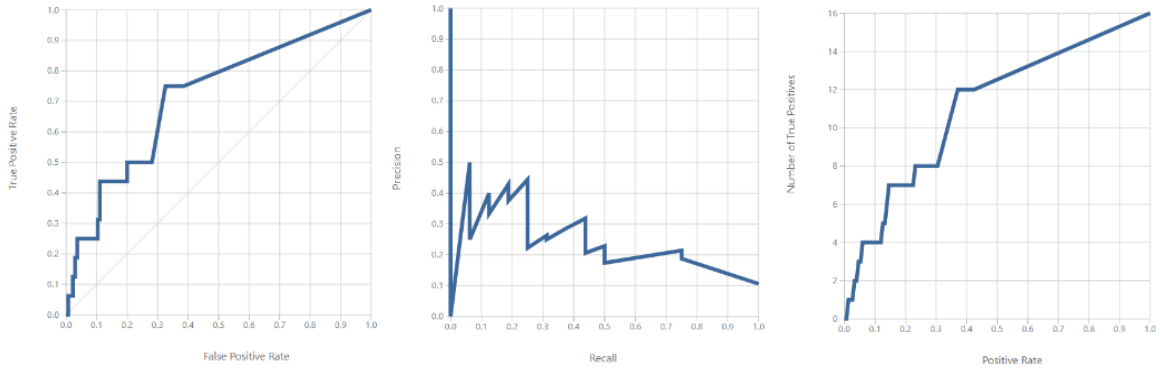
Positive Label	Negative Label
Yes	No

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	6	3	0.060	0.914	0.480	0.667	0.375	0.930	0.978	0.006
(0.800,0.900]	2	0	0.073	0.927	0.593	0.727	0.500	0.943	0.978	0.006
(0.700,0.800]	1	0	0.079	0.934	0.643	0.750	0.563	0.950	0.978	0.006
(0.600,0.700]	1	0	0.086	0.940	0.690	0.769	0.625	0.957	0.978	0.006
(0.500,0.600]	0	0	0.086	0.940	0.690	0.769	0.625	0.957	0.978	0.006
(0.400,0.500]	0	1	0.093	0.934	0.667	0.714	0.625	0.956	0.970	0.011
(0.300,0.400]	1	0	0.099	0.940	0.710	0.733	0.688	0.963	0.970	0.011
(0.200,0.300]	1	3	0.126	0.927	0.686	0.632	0.750	0.970	0.948	0.026
(0.100,0.200]	0	3	0.146	0.907	0.632	0.545	0.750	0.969	0.926	0.043
(0.000,0.100]	4	125	1.000	0.106	0.192	0.106	1.000	1.000	0.000	0.875

8.2.2 With Feature selection only



Performance metrics:



Evaluation:

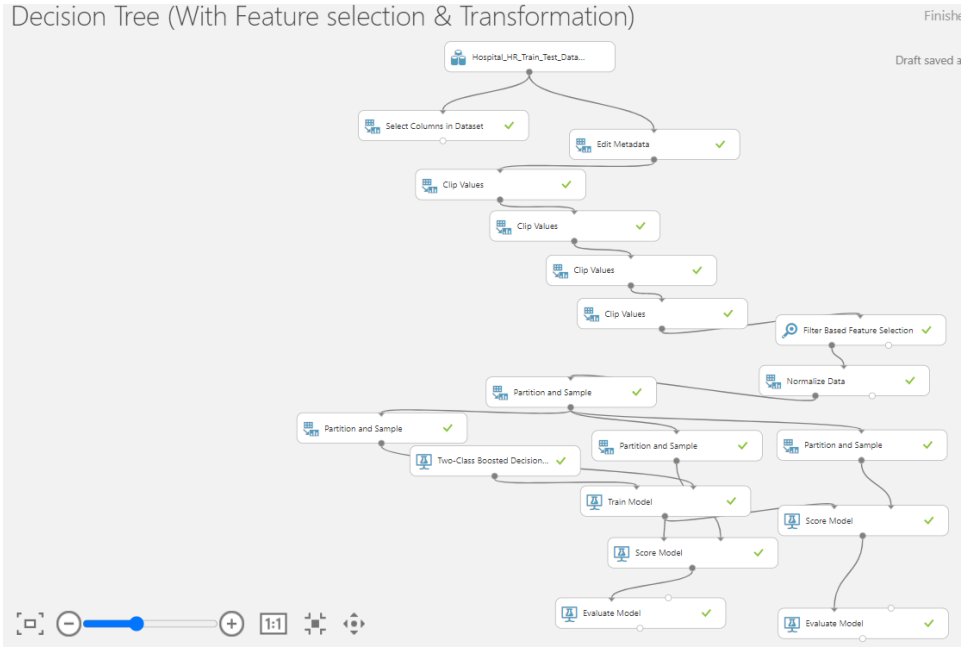
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
4	12	0.874	0.364	0.5	0.733

False Positive	True Negative	Recall	F1 Score
7	128	0.250	0.296

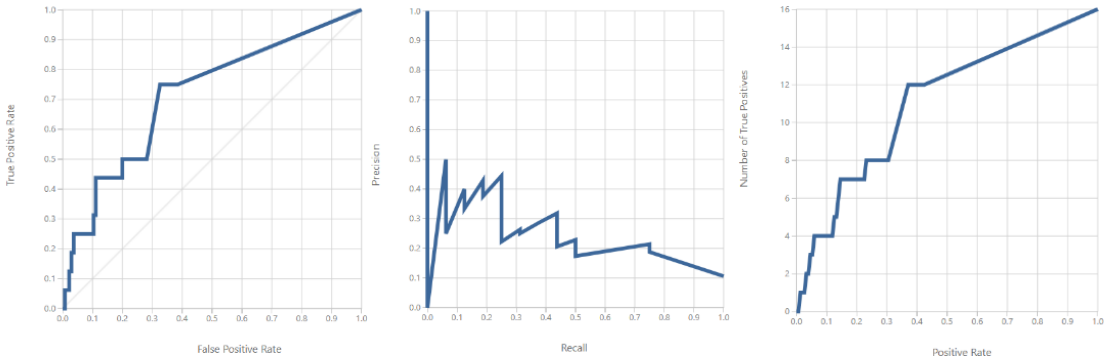
Positive Label	Negative Label
Yes	No

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	3	4	0.046	0.887	0.261	0.429	0.188	0.910	0.970	0.002
(0.800,0.900]	0	1	0.053	0.881	0.250	0.375	0.188	0.909	0.963	0.003
(0.700,0.800]	1	0	0.060	0.887	0.320	0.444	0.250	0.915	0.963	0.003
(0.600,0.700]	0	2	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.500,0.600]	0	0	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.400,0.500]	0	0	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.300,0.400]	0	1	0.079	0.868	0.286	0.333	0.250	0.914	0.941	0.009
(0.200,0.300]	0	0	0.079	0.868	0.286	0.333	0.250	0.914	0.941	0.009
(0.100,0.200]	0	2	0.093	0.854	0.267	0.286	0.250	0.912	0.926	0.012
(0.000,0.100]	12	125	1.000	0.106	0.192	0.106	1.000	1.000	0.000	0.733

8.2.3 With both Feature selection & Transformation



Performance Metrics:



Evaluation:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
4	12	0.874	0.364	0.5	0.733

False Positive	True Negative	Recall	F1 Score
7	128	0.250	0.296

Positive Label	Negative Label
Yes	No

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	3	4	0.046	0.887	0.261	0.429	0.188	0.910	0.970	0.002
(0.800,0.900]	0	1	0.053	0.881	0.250	0.375	0.188	0.909	0.963	0.003
(0.700,0.800]	1	0	0.060	0.887	0.320	0.444	0.250	0.915	0.963	0.003
(0.600,0.700]	0	2	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.500,0.600]	0	0	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.400,0.500]	0	0	0.073	0.874	0.296	0.364	0.250	0.914	0.948	0.007
(0.300,0.400]	0	1	0.079	0.868	0.286	0.333	0.250	0.914	0.941	0.009
(0.200,0.300]	0	0	0.079	0.868	0.286	0.333	0.250	0.914	0.941	0.009
(0.100,0.200]	0	2	0.093	0.854	0.267	0.286	0.250	0.912	0.926	0.012
(0.000,0.100]	12	125	1.000	0.106	0.192	0.106	1.000	1.000	0.000	0.733

8.3 Model 3: SVM

Support Vector Machines (SVM) may be used to make predictions about discrete events, such as an employee's decision to stay or go. By optimising the distance between the proposed border and the nearest points in each class, support vector machines (SVMs) attempt to identify the optimal boundary between the two groups. Support vectors are defined as the set of points that lie along the border.

The first step in developing an SVM model to forecast turnover rates is to collect information on potential motivators for employees to depart, such as job satisfaction, pay, length of service, and so on. In addition, we'd keep track of employee terminations. To further analyse this information, it would be partitioned into a training set and a test set.

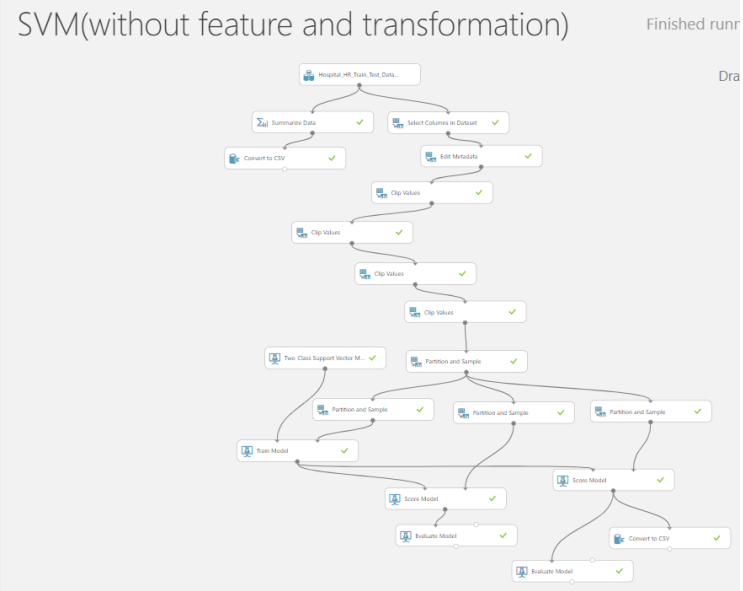
After that, we'd train the SVM algorithm on the training data to determine what boundary values best divide the two classes. Finding the boundary that maximises the margin between the support vectors from each class while reducing classification error on the training set is the goal of the procedure.

The trained SVM model may then be used to forecast whether a new hire would stay or go, depending on the values of the independent variables. Each new hire is assigned a score by the SVM model that indicates how far they are from the cutoff criteria. If the score is high,

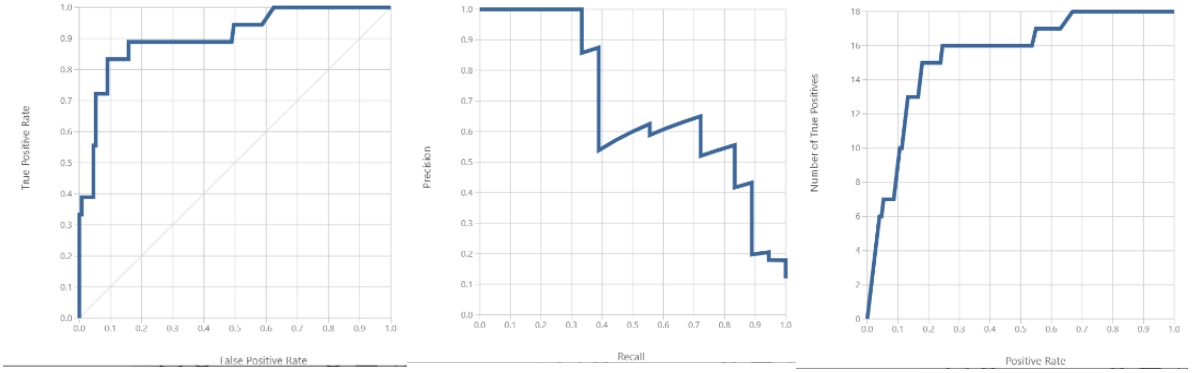
it's likely that the worker will remain with the company; otherwise, they're more likely to depart.

Accuracy, precision, recall, and F1 score are only few of the test set performance measures that may be used to assess the SVM model's efficacy. Altering hyperparameters like the kernel function, regularisation parameter, and gamma may also be used to fine-tune the model's performance.

8.3.1 Without Feature selection & Transformation



Performance Metrics:



Evaluation:

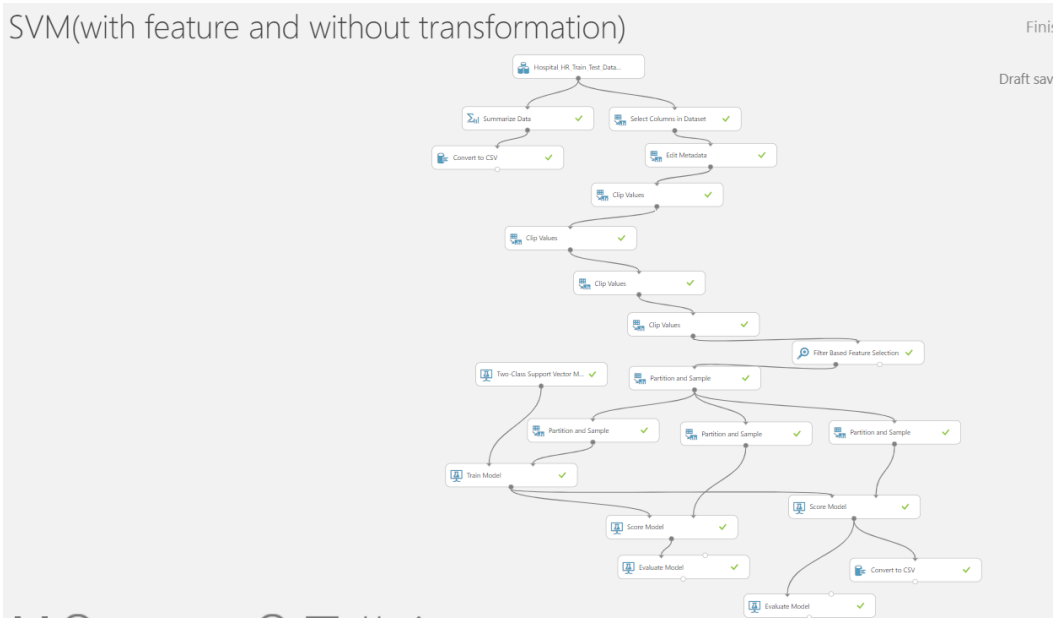
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
9	9	0.901	0.600	0.5	0.903

False Positive	True Negative	Recall	F1 Score
6	127	0.500	0.545

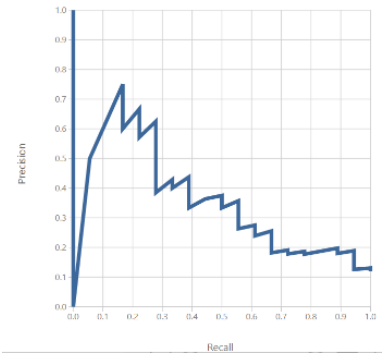
Positive Label	Negative Label
Yes	No

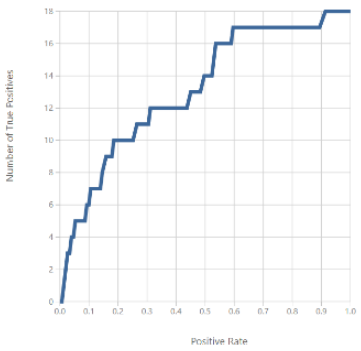
Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1	0	0.007	0.887	0.105	1.000	0.056	0.887	1.000	0.000
(0.800,0.900]	0	0	0.007	0.887	0.105	1.000	0.056	0.887	1.000	0.000
(0.700,0.800]	3	0	0.026	0.907	0.364	1.000	0.222	0.905	1.000	0.000
(0.600,0.700]	3	3	0.066	0.907	0.500	0.700	0.389	0.922	0.977	0.008
(0.500,0.600]	2	3	0.099	0.901	0.545	0.600	0.500	0.934	0.955	0.017
(0.400,0.500]	4	1	0.132	0.921	0.684	0.650	0.722	0.962	0.947	0.021
(0.300,0.400]	0	4	0.159	0.894	0.619	0.542	0.722	0.961	0.917	0.043
(0.200,0.300]	1	1	0.172	0.894	0.636	0.538	0.778	0.968	0.910	0.048
(0.100,0.200]	2	12	0.265	0.828	0.552	0.400	0.889	0.982	0.820	0.125

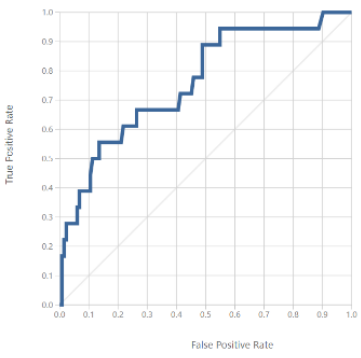
8.3.2 With Feature selection only:



Performance Metrics:







Evaluation:

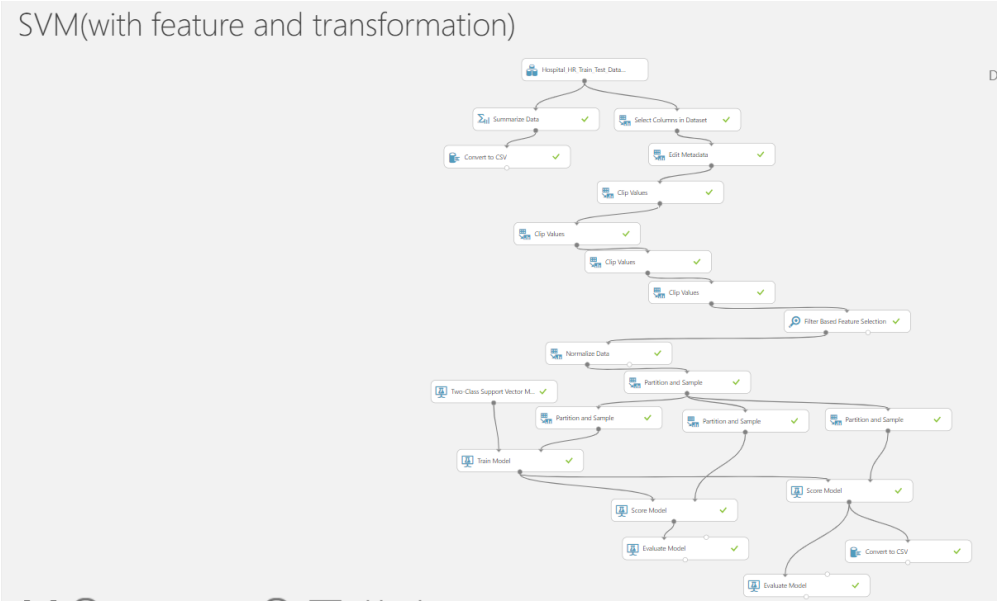
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3	15	0.894	0.750	0.5	0.761

False Positive	True Negative	Recall	F1 Score
1	132	0.167	0.273

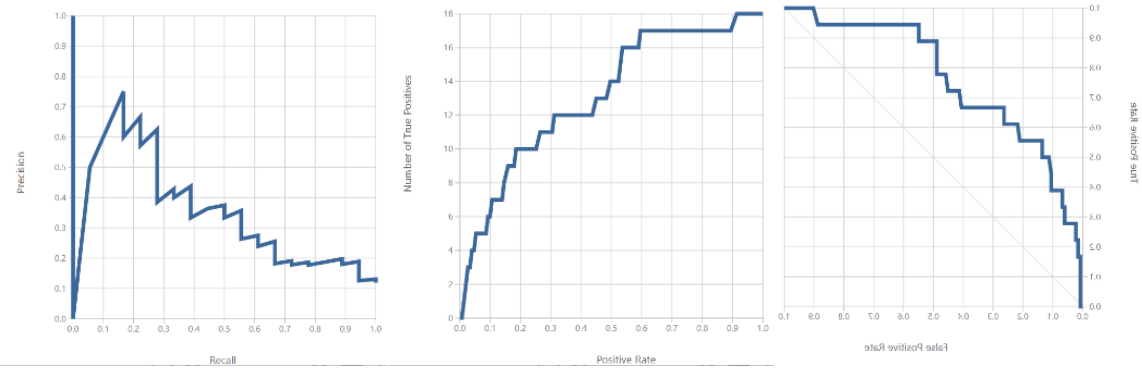
Positive Label	Negative Label
Yes	No

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.800,0.900]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.700,0.800]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.600,0.700]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.500,0.600]	3	1	0.026	0.894	0.273	0.750	0.167	0.898	0.992	0.000
(0.400,0.500]	2	2	0.053	0.894	0.385	0.625	0.278	0.909	0.977	0.003
(0.300,0.400]	0	5	0.086	0.861	0.323	0.385	0.278	0.906	0.940	0.013
(0.200,0.300]	5	10	0.185	0.828	0.435	0.357	0.556	0.935	0.865	0.046
(0.100,0.200]	2	25	0.364	0.675	0.329	0.218	0.667	0.938	0.677	0.159

8.3.3 With both Feature selection and Transformation



Performance metrics:



Evaluation:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3	15	0.894	0.750	0.5	0.761

False Positive	True Negative	Recall	F1 Score
1	132	0.167	0.273

Positive Label	Negative Label
Yes	No

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.800,0.900]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.700,0.800]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.600,0.700]	0	0	0.000	0.881	0.000	1.000	0.000	0.881	1.000	0.000
(0.500,0.600]	3	1	0.026	0.894	0.273	0.750	0.167	0.898	0.992	0.000
(0.400,0.500]	2	2	0.053	0.894	0.385	0.625	0.278	0.909	0.977	0.003
(0.300,0.400]	0	5	0.086	0.861	0.323	0.385	0.278	0.906	0.940	0.013
(0.200,0.300]	5	10	0.185	0.828	0.435	0.357	0.556	0.935	0.865	0.046
(0.100,0.200]	2	25	0.364	0.675	0.329	0.218	0.667	0.938	0.677	0.159

9 Results

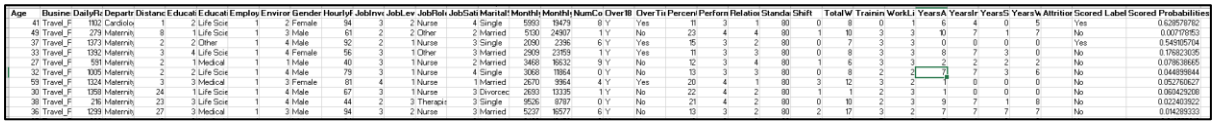
e

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
27	26	0.932	0.844	0.5	0.933

False Positive	True Negative	Recall	F1 Score
5	395	0.509	0.635

Positive Label	Negative Label
Yes	No

The Logistic regression model without feature selection and transformation gave the best AUC score and accuracy compared to other models. Hence, it is the best fit for the dataset. A new dataset was used to validate the model. Below are the results the prediction,



10 Conclusion

The three models applied show their best results in the first variation where no features are selected, and no transformation methods applied.

After observing the evaluation parameters – accuracy, precision and recall, we can conclude that **Logistic Regression** is the most accurate and suitable model for this dataset.