Employee Retention Program - Recruit and Retain Talent Team – Executive Summary

Employee retention refers to keeping employees in organizations and preventing them from leaving at any cost. Attracting and retaining the best employees is a stumbling block for any company. There are a variety of reasons why an employee may resign or attempt to leave their work, some of them could be:

1. Many job prospects outside the company
2. Low payment tiers/monthly salaries are a source of dissatisfaction.
3. Workplace adaptability
4. Uncomfortable/Unbalance work life
5. There are no employment perks, such as leave policies.

Employees are the most vital component of any business. They have an associate degree and are an essential element of the business; without them, the company cannot envisage growing and cannot also be ineffective in achieving any one of the organization's goals.

Any organization does not have the authority to terminate any of its employees unless it has a strategy or a plan in place to keep the employees. Different companies have different methods of retaining employees, but what matters is the company's plan for retaining employees.

The HR department from ABC IT company that Recruit and Retain jobs ensures that its employees don’t quit their jobs. They want to compensate employees with retention bonus and provide employment perks to achieve maximum retention.

Problem Statement

The ABC IT company's HR department, which recruits and retains employees, has limited funds; they cannot compensate every employee and are concerned about who they should incentivize to increase retention rates. They require the services of a data scientist to gain insight into employees who are leaving the company and determine who should be compensated with retention bonus.

Problem Solution

Using a data set containing 4653 observations containing employee details and their decisions to quit or stay with the organization, we "Team 8" data scientists will generate predictions on the data set and provide data insights and recommendations on whom they should reward.

Data Description:

The data set has 4635 observations with 9 different variables as explained below,

Independent Variables:

- Education: Education Level - Bachelors, Masters and PHD.
- Joining Year: Year of Joining Company.
- City: City office where Posted: Bangalore, Pune, and New Delhi.
- Payment Tier: Payment Tier 1- Highest, Payment Tier 2 – Medium, Payment Tier 3 – Lowest.
- Age: Current Age
- Gender: Gender of Employee
- Ever Benched: Ever Kept out of Project For more than one month or more.
- Experience: Experience in Current Field.

Dependent (Target) Variable:

- Leave (1) or Not (0)

The dataset contains 3 binary, 4 categorical and 2 continuous variables.

More about Data:

The following images explains the Variable summary and Information about any missing values in Class, Interval and Target Variables in detail.

Variable Summary:

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	2
INPUT	INTERVAL	3
INPUT	NOMINAL	3
TARGET	BINARY	1

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
LeaveOrNot	TARGET	2

The below images explain that the Class, Interval and Target Variables do not have any missing values.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City	INPUT	3	0	Bangalore	47.88	Pune	27.25
TRAIN	Education	INPUT	3	0	Bachelors	77.39	Masters	18.76
TRAIN	EverBenched	INPUT	2	0	No	89.73	Yes	10.27
TRAIN	Gender	INPUT	2	0	Male	59.70	Female	40.30
TRAIN	PaymentTier	INPUT	3	0	3	75.05	2	19.73
TRAIN	LeaveOrNot	TARGET	2	0	0	65.61	1	34.39

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	29.39329	4.826087	4653	0	22	28	41	0.905195	-0.29982
ExperienceInCurrentDomain	INPUT	2.905652	1.55824	4653	0	0	3	7	-0.16256	-0.96941
JoiningYear	INPUT	2015.063	1.863377	4653	0	2012	2015	2018	-0.11346	-1.24232

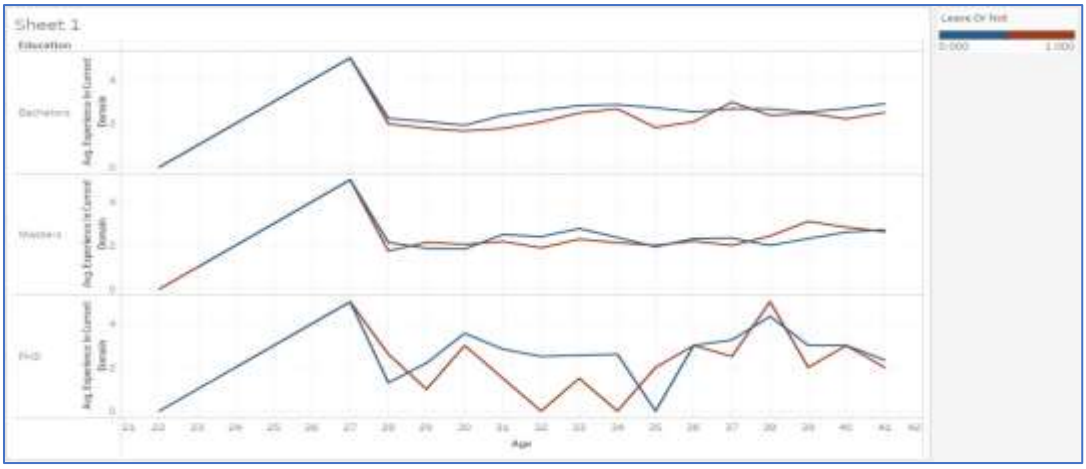
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	LeaveOrNot	TARGET	0	3053	65.6136
TRAIN	LeaveOrNot	TARGET	1	1600	34.3864

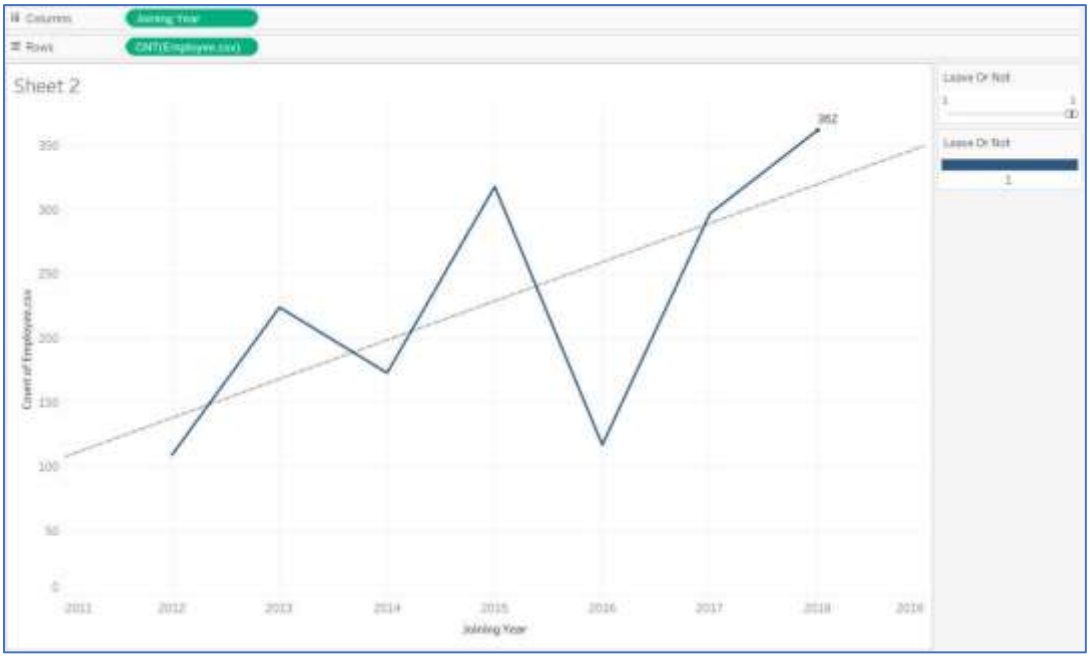
Data Visualization:

Initial Results Observed: There are some outliners that have been observed while exploring the data. There is no linear correlation between age and experience. In the first line graph (Bachelors) at age 37 we can see that average experience of people who have never left the company is less than average experience of the people who left the company. Same behaviors (outliers) are observed for master's and PHD employee's as well. Refer to the below graph.

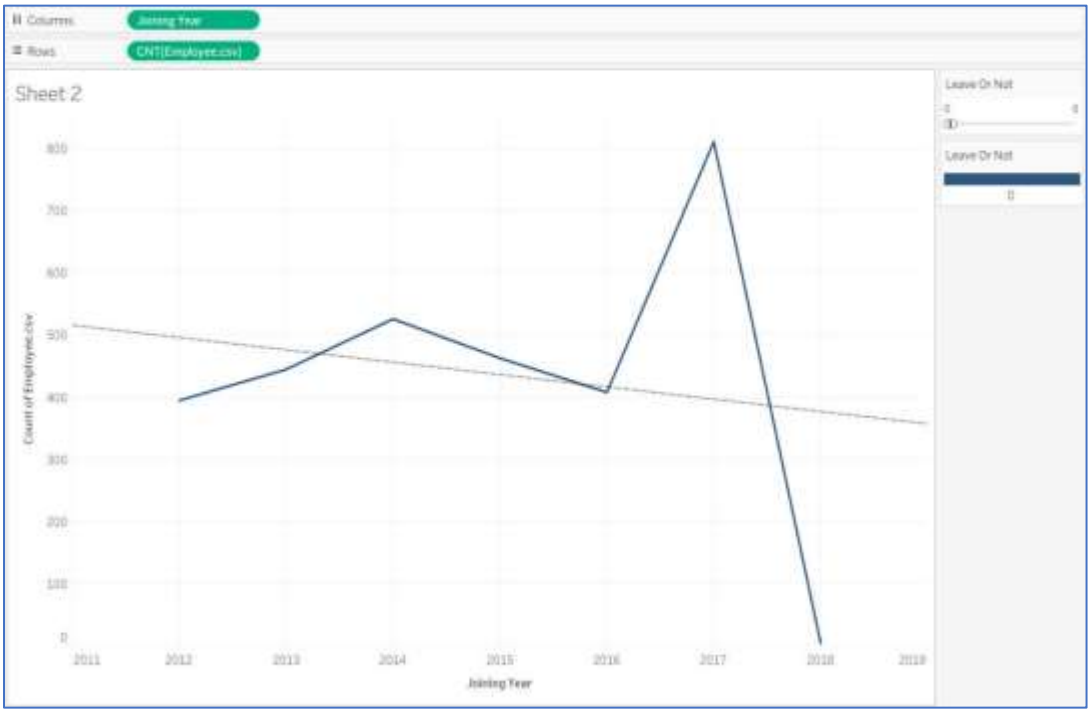


Trend Lines:

We can see a trend that employees leaving the company are increasing in the coming years. Refer to the below graph

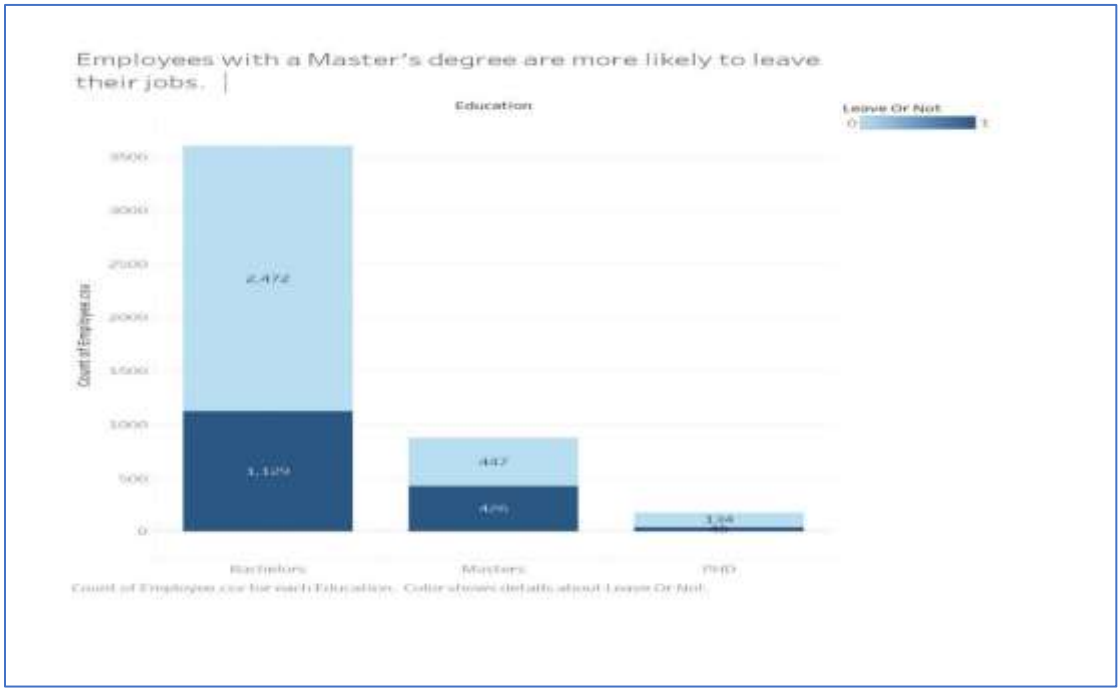


We can see a trend that employees staying with the company is very less in the coming years. Refer to the below graph.

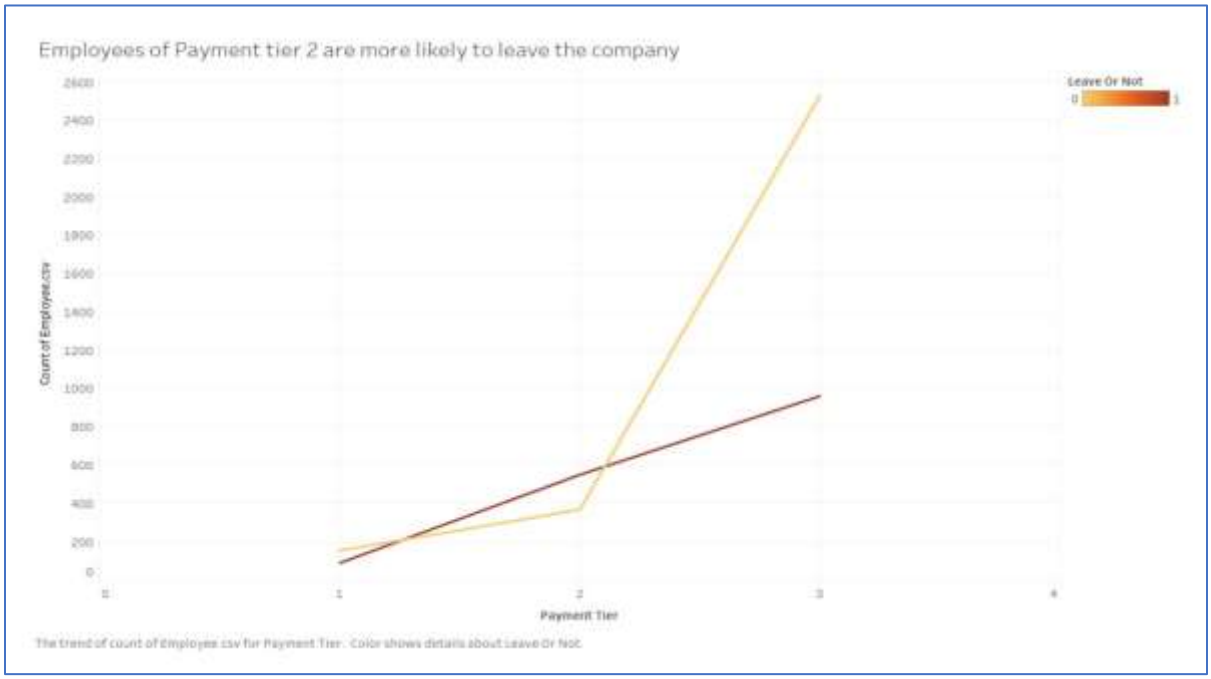


Additional Graphs:

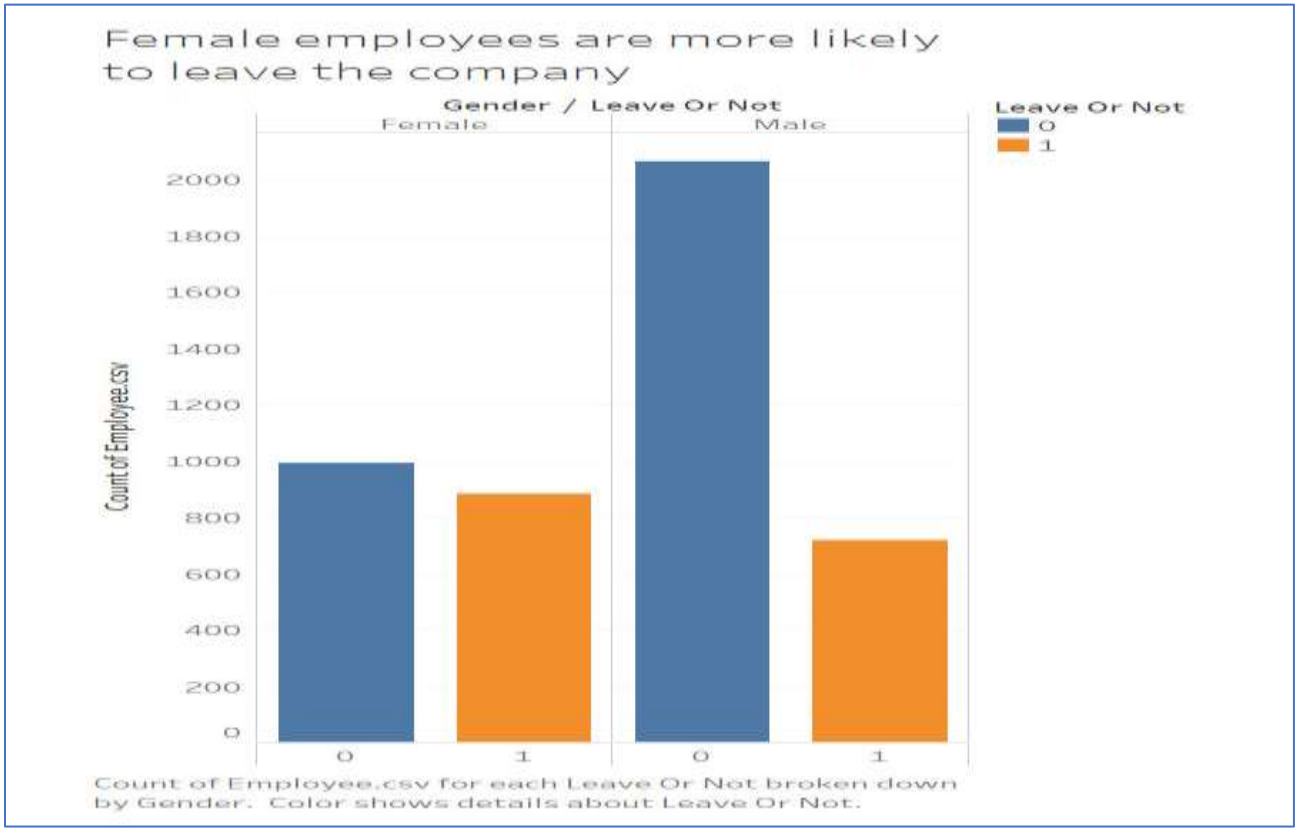
The graph shows that nearly half of Masters' degree holding employees are likely to leave the company.



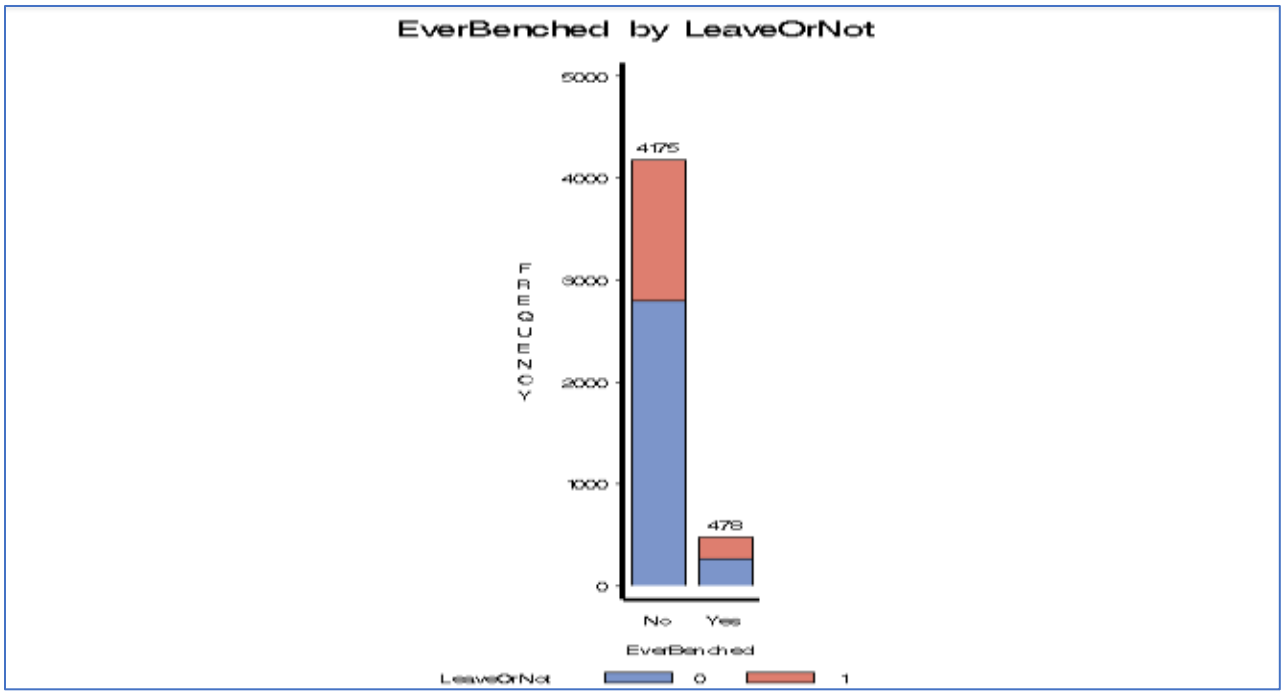
The graph shows that employees of Payment Tier 2 are likely to leave the company.



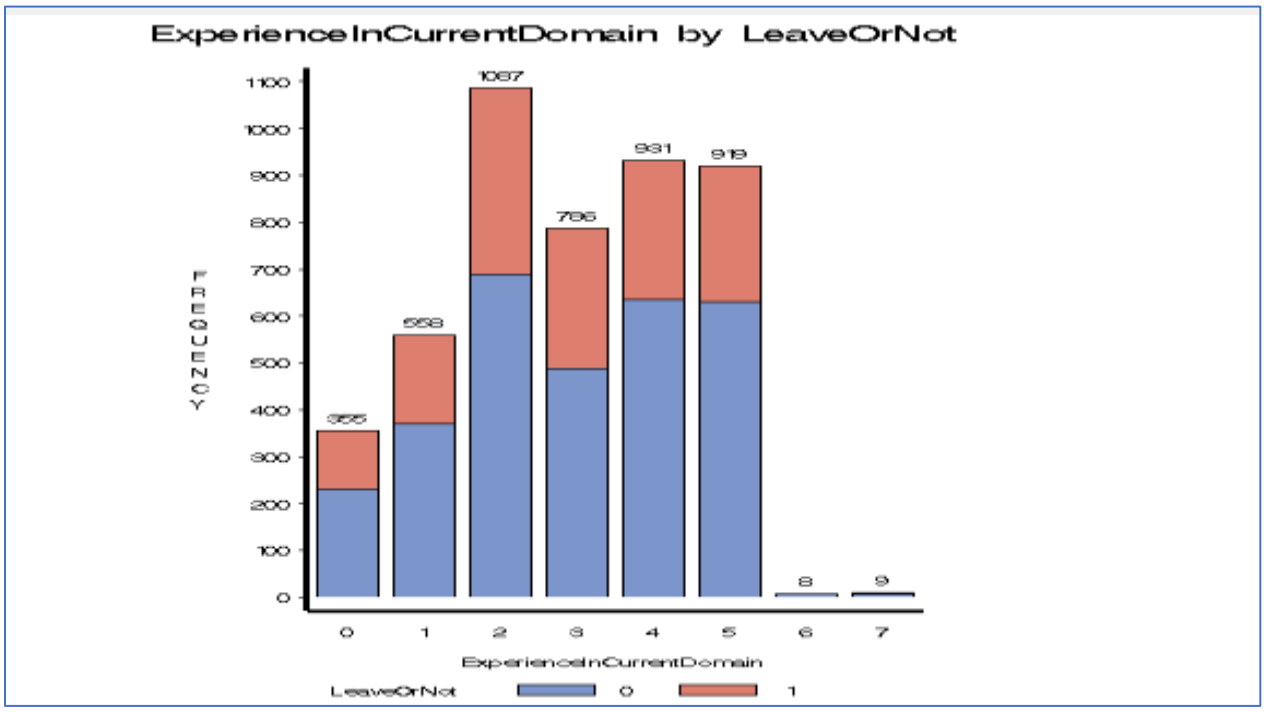
The graph shows that high number of Female employees are likely to leave the company



The graph shows the values if employees who are ever benched by Leave or Not



The below graph shows the employees who are more likely to leave or stay with the company based on their work experience.



Visualization Findings:

Female employees and employees with a master's degree are more likely to leave the organization, as seen in the above visualization graphs. Payment tier is one of the most common reasons for employees to leave the organization. Employees on the second category of salary are more likely to leave. We also notice that based on Ever benched and job experience, there isn't much of an impact on employees leaving.

We think the employees are leaving due of the poor pay scales. We also assume that the female employees are leaving since they are not receiving any employee benefits.

Using modeling techniques, we will now predict the employees who need to be compensated so that they do not leave the organization. We'd also like to provide additional information to management based on our findings.

Prediction Models and Findings:

We have used Linear regression, Logistic regression, and Decision tree models to Train, validate and test the data.

1. We take the data and divide it into three sections: 70%, 20%, and 10%.
2. We set aside 10% of the data. At the end, this will be used as new data to make predictions.
3. Use 70% of the data to train the model and 20% to validate the model.

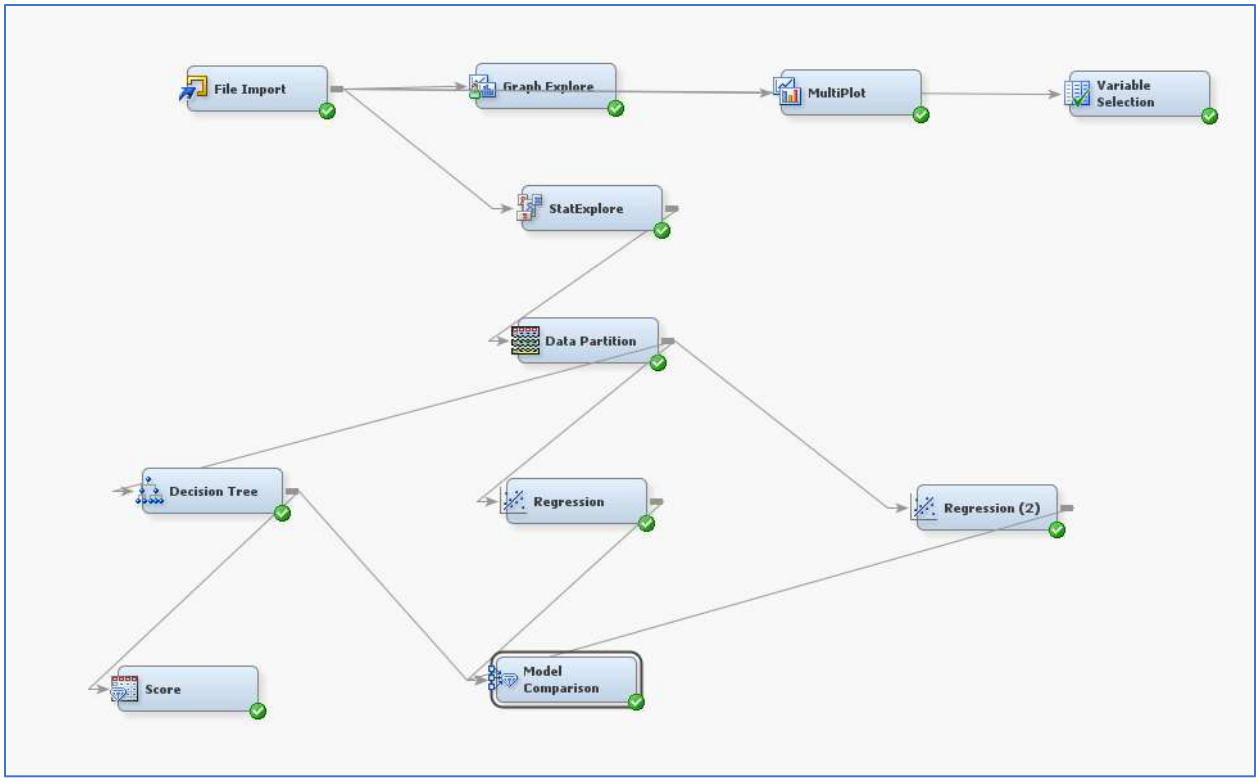
Data Partition:

As shown below we have set 70% to Train, 20% to Validate and 10% to Test.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Stat_TRAIN	4653
TRAIN	EMWS1.Part_TRAIN	3256
VALIDATE	EMWS1.Part_VALIDATE	930
TEST	EMWS1.Part_TEST	467

Diagram Flow in SAS Tool:



The results of each model are as follows,

Linear Regression results:

Analysis of Maximum Likelihood Estimates

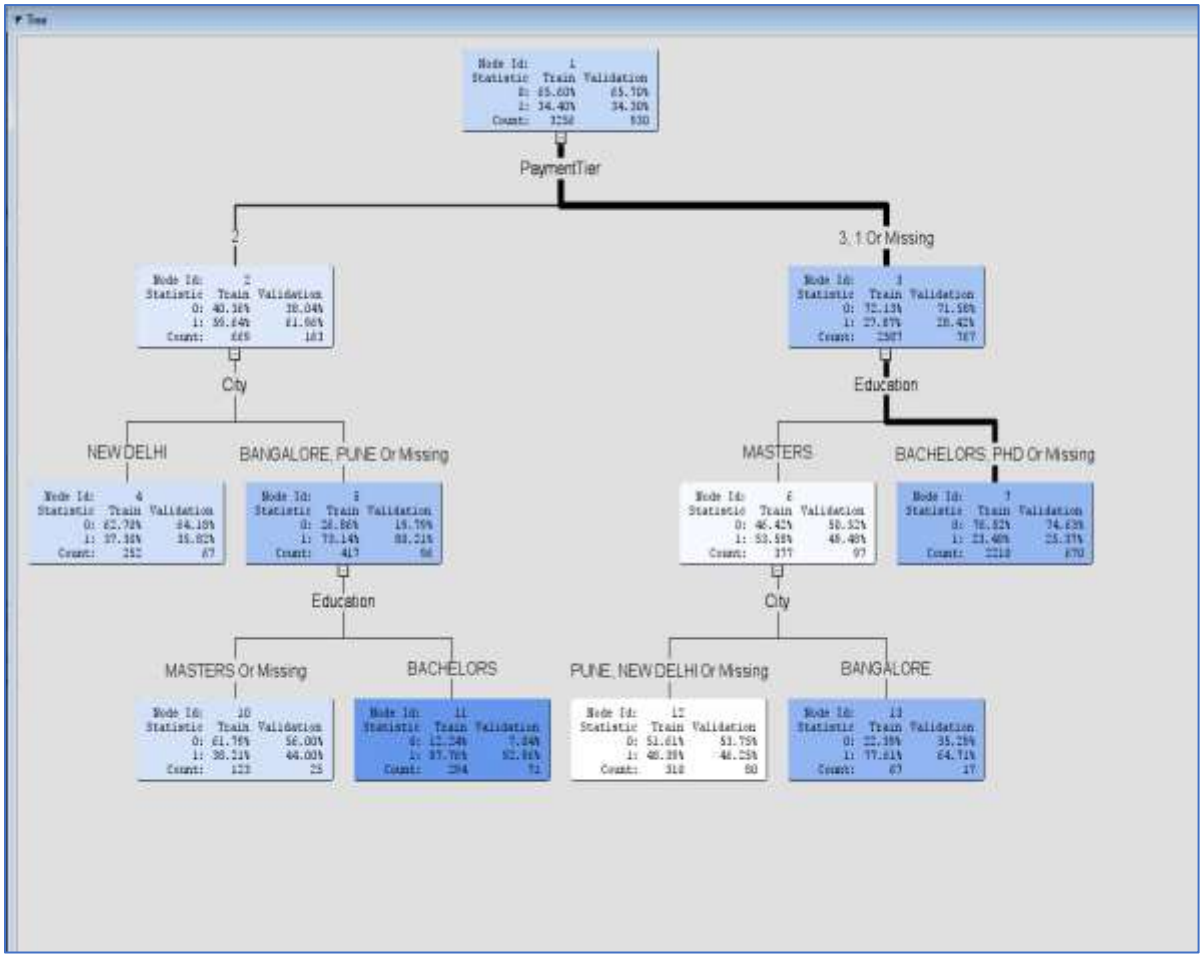
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	0.3277	0.0558	5.87	<.0001
Age		1	0.00631	0.00164	3.85	0.0001
City	Bangalore	1	-0.0112	0.0119	-0.94	0.3459
City	New Delhi	1	0.1218	0.0133	9.19	<.0001
Education	Bachelors	1	0.0792	0.0167	4.73	<.0001
Education	Masters	1	-0.1043	0.0185	-5.64	<.0001
EverBenched	No	1	0.0485	0.0128	3.80	0.0001
ExperienceInCurrentDomain		1	0.00931	0.00504	1.85	0.0650
Gender	Female	1	-0.0902	0.00831	-10.85	<.0001
PaymentTier	1	1	0.0715	0.0241	2.96	0.0031
PaymentTier	2	1	-0.1375	0.0175	-7.85	<.0001

Logistic Regression results:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	0.9547	0.2928	10.63	0.0011		2.598
Age		1	-0.0335	0.00868	14.95	0.0001	-0.0881	0.967
City	Bangalore	1	0.0616	0.0618	0.99	0.3190		1.064
City	New Delhi	1	-0.5877	0.0701	70.35	<.0001		0.556
Education	Bachelors	1	-0.3752	0.0876	18.35	<.0001		0.687
Education	Masters	1	0.5111	0.0955	28.64	<.0001		1.667
EverBenched	No	1	-0.2406	0.0637	14.28	0.0002		0.786
ExperienceInCurrentDomain		1	-0.0493	0.0262	3.55	0.0597	-0.0420	0.952
Gender	Female	1	0.4369	0.0420	108.44	<.0001		1.548
PaymentTier	1	1	-0.3232	0.1215	7.08	0.0078		0.724
PaymentTier	2	1	0.6104	0.0863	50.05	<.0001		1.841

Decision Tress: Build Decision tree with high significant variables.



About Results (β values):

The outputs in the above images of each model shows the coefficient's (β values). Payment Tier, Gender, City, and Education, all have a Pr/Chi-sqr of <0.001, indicating that they are highly significant variables. Also, these parameters could have a positive or negative impact on employees leaving the company.

Reasons to Choose Logistic regression as our model:

Logistic Regression assumes that the data is linearly (or curvy linearly) separable in space on to exactly two planes.

Decision Trees are non-linear classifiers; they do not require data to be linearly separable. They Bisect the sample space in to smaller and smaller Regions

We are sure that our data set divides in to exactly two separable parts, so we have chosen to go with Regression model as it is performing best on our data.

We have chosen Logistic regression over Linear regression because of the following reasons,

1. The dependent variable (Target Variable) in our dataset is binary and Logistic regression is highly used in this case.
2. Also, the Misclassification rates and Average squared errors for Logistic regression are slightly less when compared to Linear regression. Please see the image on the left side of this page.
3. The event classification table for both models are given below on the right-side of the page, as you can see the False Negative values are better for Logistic regression.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg	Logistic Regression	0.26882	0.19218	0.27611	0.18945
	Reg2	Linear Regression	0.27527	0.19374	0.28409	0.19142

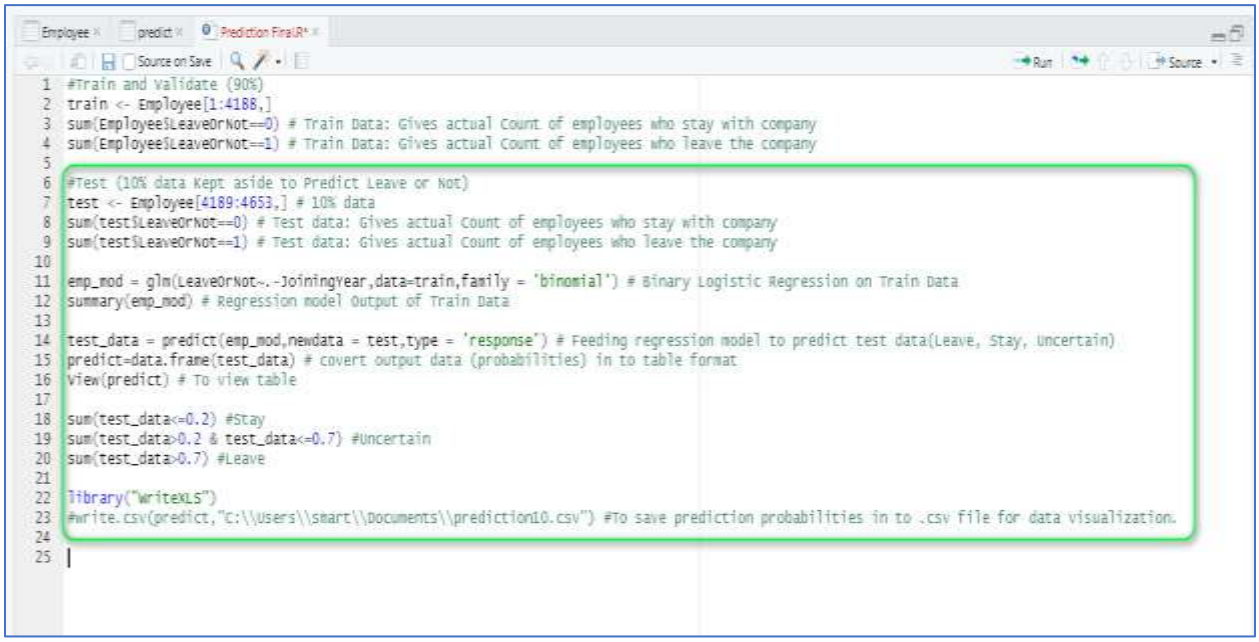
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg2	Linear Regression	TRAIN	LeaveOrNot		752	1963	173	368
Reg2	Linear Regression	VALIDATE	LeaveOrNot		220	575	36	99
Reg	Logistic Regression	TRAIN	LeaveOrNot		715	1952	184	405
Reg	Logistic Regression	VALIDATE	LeaveOrNot		214	575	36	105

Prediction (ŷ):

we have used R studio to run Logistic regression model on the 10% data (considered as new data) that was kept aside at the beginning of the data partition. There are 465 observations in this 10% data set.

The R studio code is shown in the figure below. Each line's explanation is provided in terms of comments. The code for predicting the probabilities of each observation and exporting the resultant data to an Excel file is highlighted in green.



Why Probabilities?

Before we look at the prediction outcomes, we need to discuss how these outputs can help enhance retention rates.

We know that the probabilities of an employee quitting the organization range from 0 to 1. We also know that the probability number closest to 0 indicates that the employee has a low interest in leaving the company, while the probability value closest to 1 indicates that the person has a great interest in leaving the organization. Based on this, we can say that employees with probability values ranging from 0.3 to 0.7 are most likely unsure about whether to leave the organization.

Overall, the probability of an employee leaving the organization is divided into three categories:

- 0 % to 20% (0 to 0.2) - indicates that the probability of employees leaving the company is very low. There is no need to Incentivize because they won't leave the company anyway.

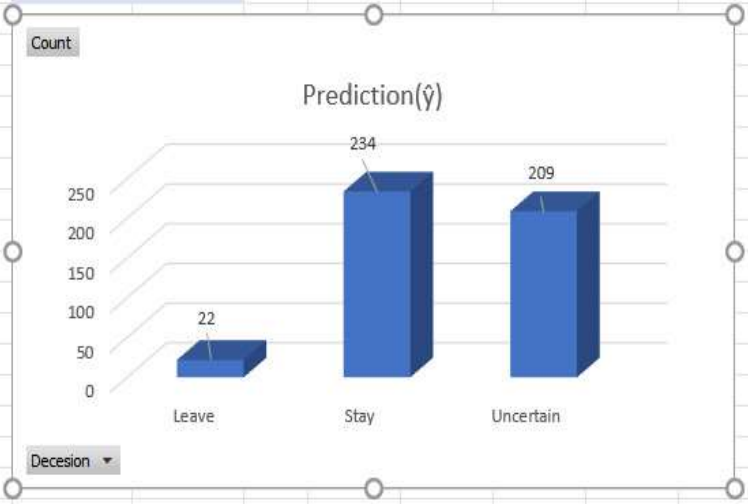
- 70% to 100% (0.7 to 1) indicates that the probability of employees leaving the organization is very high. There is no need to incentivize because they will leave the company anyway, even if they are incentivized.
- 20% to 70% (0.2 to 0.7) - the probability of employees who are uncertain (Undecided) whether to leave or stay with the company. These employees need to be incentivized to obtain maximum retention.

Results:

The following picture shows the probabilities of each observation in excel sheet, and bar chart that shows count of employees that are leaving, staying and uncertain in the company.

Test_Observations	Probability	Decesion
4221	0.581429898	Uncertain
4222	0.221155617	Stay
4223	0.117732334	Stay
4224	0.768634445	Leave
4225	0.358915988	Uncertain
4226	0.231398209	Stay
4227	0.806980715	Leave
4228	0.148342024	Stay
4229	0.340450617	Uncertain
4230	0.194121188	Stay
4231	0.278679987	Stay
4232	0.110227337	Stay
4233	0.222727673	Stay
4234	0.46239642	Uncertain
4235	0.367370611	Uncertain
4236	0.318720945	Uncertain
4237	0.156794143	Stay
4238	0.304791511	Uncertain
4239	0.519720779	Uncertain
4240	0.2667256	Stay
4241	0.457075374	Uncertain
4242	0.77613803	Leave
4243	0.376601682	Uncertain
4244	0.47542477	Uncertain
4245	0.192401635	Stay
4246	0.244024448	Stay
4247	0.378437997	Uncertain
4248	0.369279011	Uncertain
4249	0.394992897	Uncertain
4250	0.463522271	Uncertain
4251	0.163742572	Stay
4252	0.322457408	Uncertain
4253	0.577132209	Uncertain

Decesion	Count
Leave	22
Stay	234
Uncertain	209
Grand Total	**465**

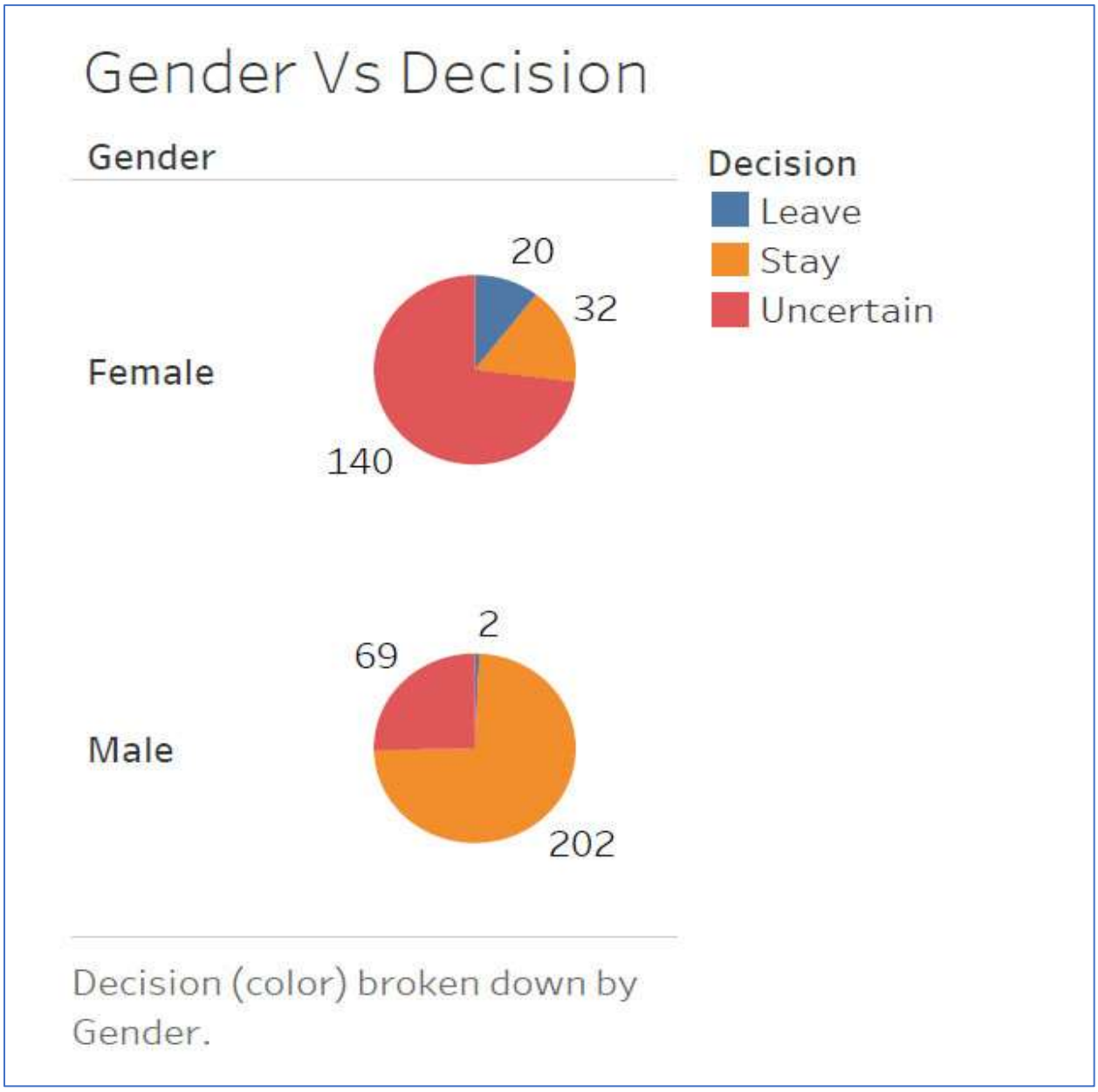


22 employees are leaving, 234 are staying, and 209 are unsure. There may be more employees i n the unsure group who will leave the company. To increase retention, these employee issues must be addressed.

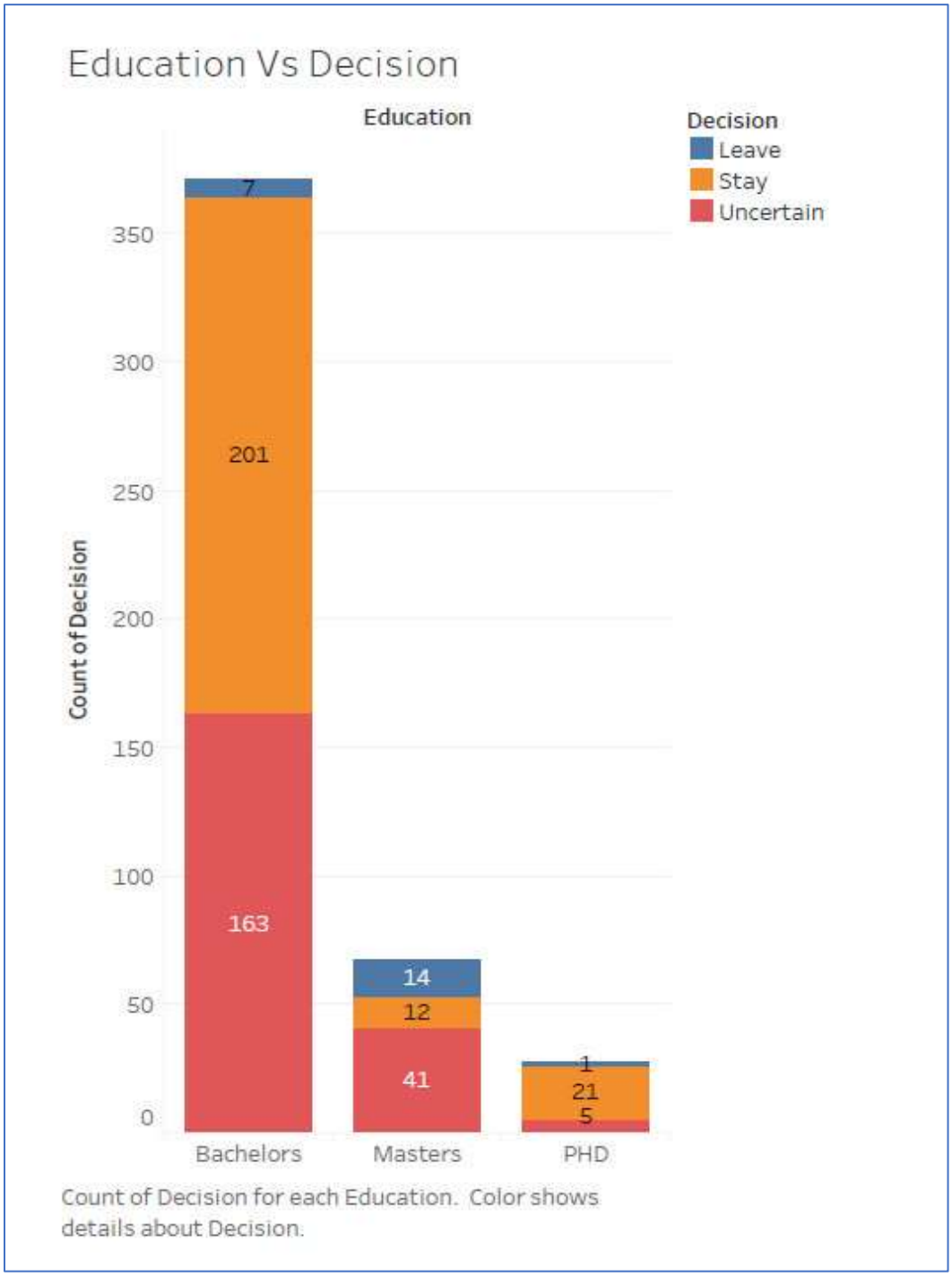
Managerial or Policy Implications/ Recommendations:

we used visualization to gain further insights from the prediction results and the following are the surprising insights.

The Pie Chart depicts that the Female employees are more likely to leave the company compared to Male employees. It is very unlikely to see that almost 90% of female employee are in Uncertain and leave category and only 10% are likely to stay in the company.



The Bar graph shows that the employees with master's degree are more likely to leave company compared to others.



Conclusion:

We can conclude from the data visualization graphs and β values that payment tier is one of the most prevalent reasons why employees leave the company. Offering an employee retention bonus is the ideal solution, but due to company's limited budget and it is not a not a good idea to distribute budget equally to every employee. So, employees in the 0 to 20% (0 to 0.2) category do not need to be incentivized because they will not leave the company anyway, whereas employees in the 70% to 100% (0.7 to 1) category do not need to be incentivized because they will leave the company regardless of whether they are incentivized. The highest retention rate will be achieved by compensating 20% to 70% (0.2 to 0.7) of category employees.

Also, 90% of Female employees are unsure or leaving the company, management must develop a Female Employee Benefits policy to prevent female employees from quitting the organization.

By utilizing modeling and visualization tools as a data scientist, we are pleased that we were able to explain the common factors for employees leaving the company, provide advice on how to best allocate their limited budget, which categories of employees the employer should reward, additional information on which types of employees are more likely to leave, and policy suggestions. We are confident that these results will ensure that the organization's retention rate is as high as possible.