**Predicting Employee Retention**

# **Problem Statement**

A mid-sized technology company is facing challenges with high employee turnover, which

disrupts project workflows and increases recruitment costs. The HR department recognizes that traditional exit interviews and surveys are not sufficient to address this issue, as they are

reactive measures that occur after an employee has already decided to leave.

The company seeks to adopt a proactive, data-driven approach to predict which employees are likely to leave. By leveraging historical employee data, the organization aims to identify patterns and risk factors associated with attrition. This will enable targeted interventions to retain valuable employees and maintain workforce stability.

# **Business Objective**

Develop a logistic regression model to predict the likelihood of employee retention based on

available data, including demographic details, performance metrics, survey sentiment scores, and workload-related factors, enabling the HR department to implement timely retention strategies and reduce turnover rates.

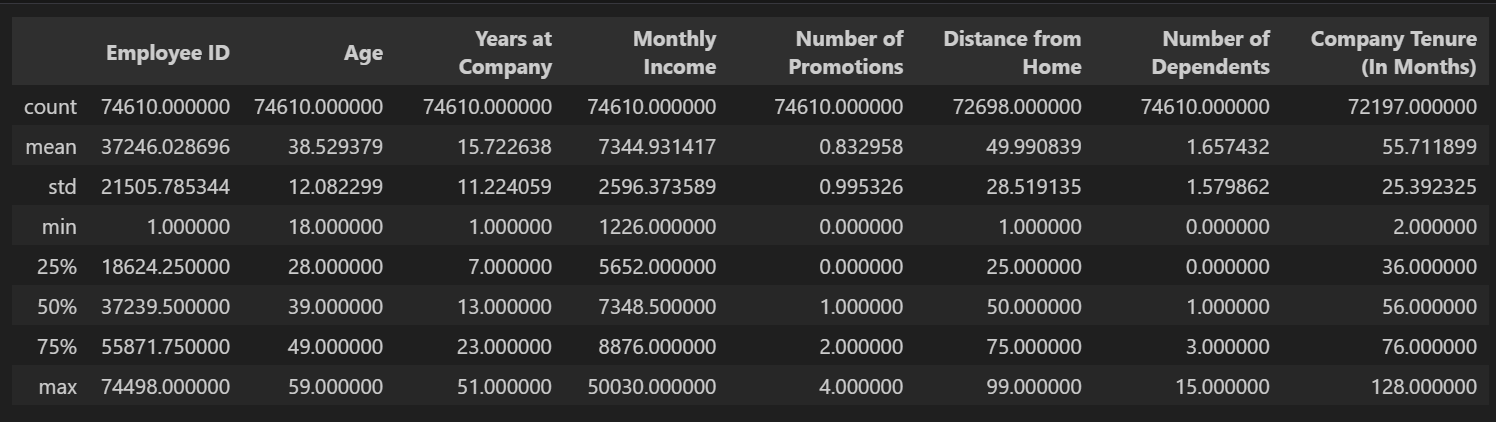
# **Approach**

1. **Data Preparation**

A csv file of 74,610 rows containing data related to employees with their details related to demography and job history was provided. Following are the columns present in the data:

* **Employee ID:** A unique identifier assigned to each employee.
* **Age:** The age of the employee, ranging from 18 to 60 years.
* **Gender:** The gender of the employee
* **Years at Company:** The number of years the employee has been working at the company.
* **Monthly Income:** The monthly salary of the employee, in dollars.
* **Job Role:** The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media.
* **Work-Life Balance:** The employee's perceived balance between work and personal life, (Poor, Below Average, Good, Excellent)
* **Job Satisfaction:** The employee's satisfaction with their job: (Very Low, Low, Medium, High)
* **Performance Rating:** The employee's performance rating: (Low, Below Average, Average, High)
* **Number of Promotions:** The total number of promotions the employee has received.
* **Distance from Home:** The distance between the employee's home and workplace, in miles.
* **Education Level:** The highest education level attained by the employee: (High School, Associate Degree, Bachelor’s Degree, Master’s Degree, PhD)
* **Marital Status:** The marital status of the employee: (Divorced, Married, Single)
* **Job Level:** The job level of the employee: (Entry, Mid, Senior)
* **Company Size:** The size of the company the employee works for: (Small, Medium, Large)
* **Company Tenure:** The total number of years the employee has been working in the industry.
* **Remote Work:** Whether the employee works remotely: (Yes or No)
* **Leadership Opportunities:** Whether the employee has leadership opportunities: (Yes or No)
* **Innovation Opportunities:** Whether the employee has opportunities for innovation: (Yes or No)
* **Company Reputation:** The employee's perception of the company's reputation: (Very Poor, Poor, Good, Excellent)
* **Employee Recognition:** The level of recognition the employee receives:(Very Low, Low, Medium, High)
* **Attrition:** Whether the employee has left the company, encoded as 0 (stayed) and 1 (Left).

Following is the basic statistical summary of the numerical columns of dataset:



## **Data Cleaning**

* 1. **Handling Missing Data:** Out of all the columns present in the dataset, following columns had missing data
     1. Distance from Home
     2. Company Tenure (In Months)

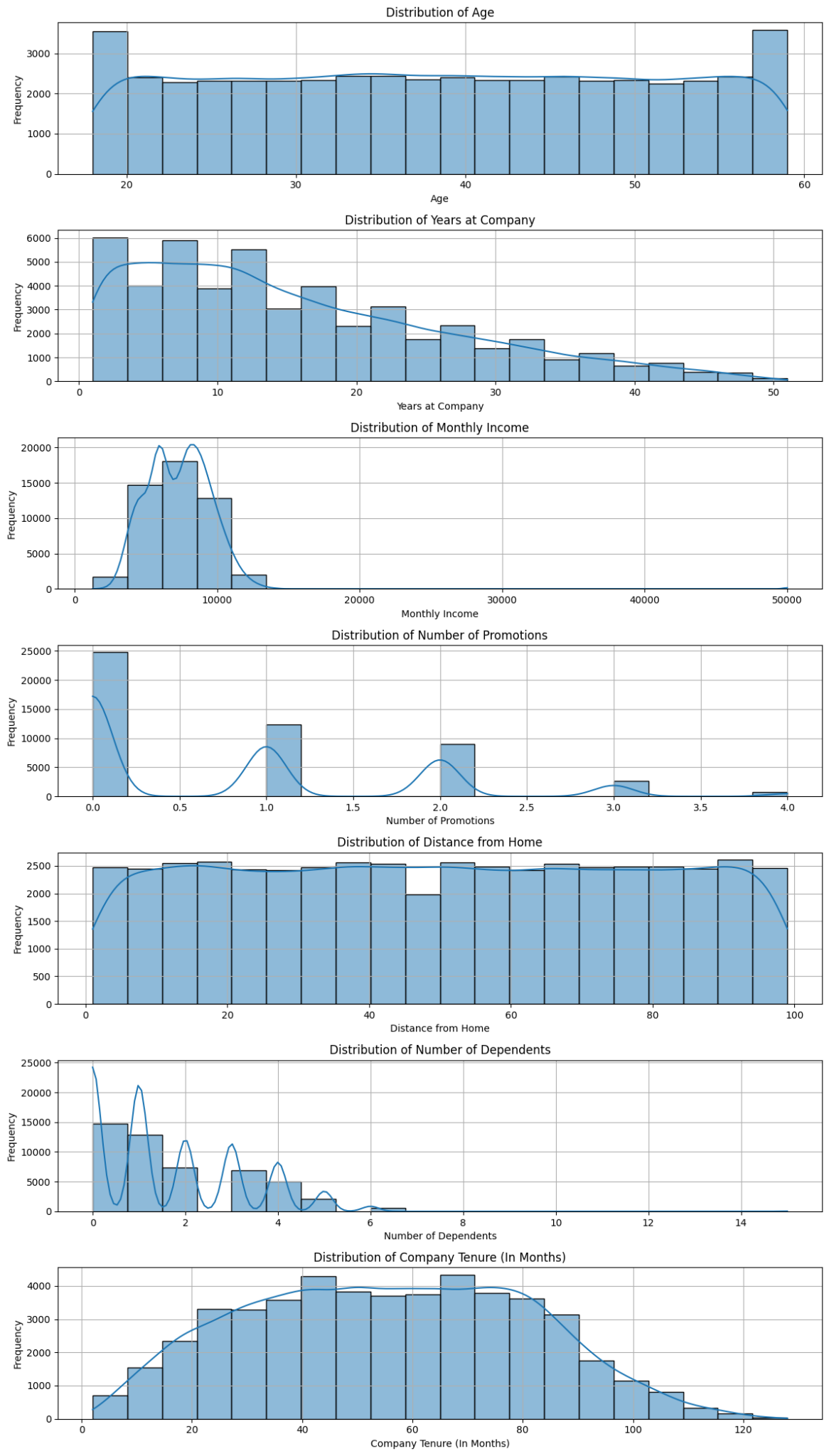
Since the number of rows having missing data was less than 5% of total rows, the missing rows were dropped. Row count after null values were removed was 70635, which is 94.6% of the original dataset.

* 1. **Checking Redundant Values in Categorical Columns:** Unique values along with value counts were checked for all columns in order to drop any redundant value if present. No such values were found in data.
  2. **Remove Unnecessary Columns:** “Employee ID” column was removed as it was not helpful in the process of model creation and prediction.

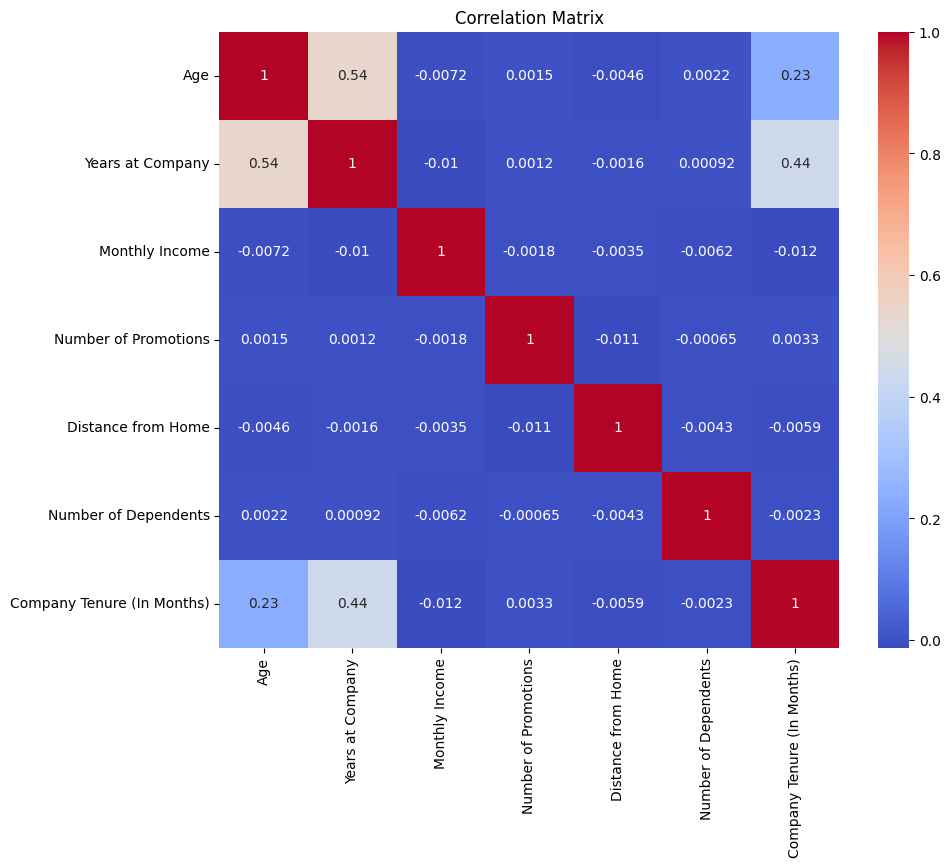
1. **Training and Validation Split**

Data was split into training and validation set in the ratio of 70:30. This was done using the train\_test\_split method.

1. **EDA on Training Set**
2. **Univariate Analysis:** Distribution of numerical columns were checked for the training data. Following are the distributions of all the numerical columns:

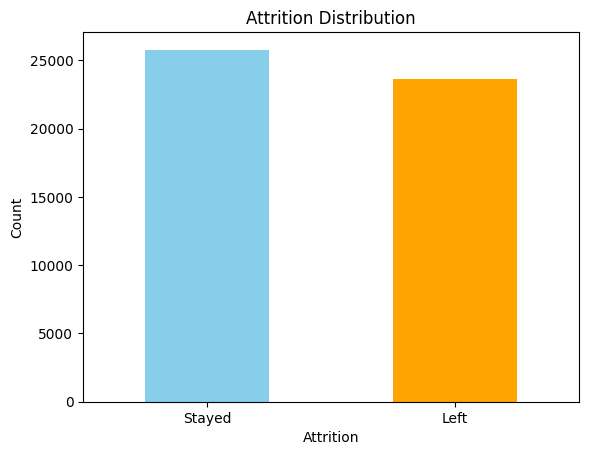


1. **Collinearity:** Multicollinearity was checked among the numerical columns present in data.



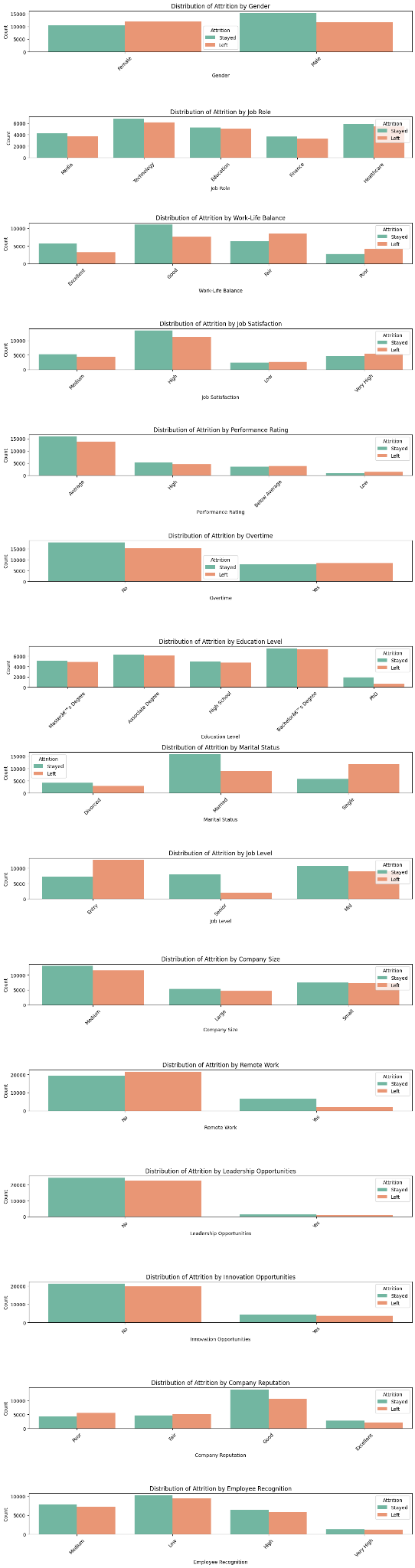
We see that there is collinearity present between columns Age and Years at Company, and also between Age and Company Tenure. This was later handled during feature selection, where Years at Company was taken as a significant feature, and Age and Company Tenure were found to be not that significant through RFE.

1. **Data Balance:** Distribution for Attrition column was plotted in order to understand out of the total training universe, how many people stayed and how many left the company.

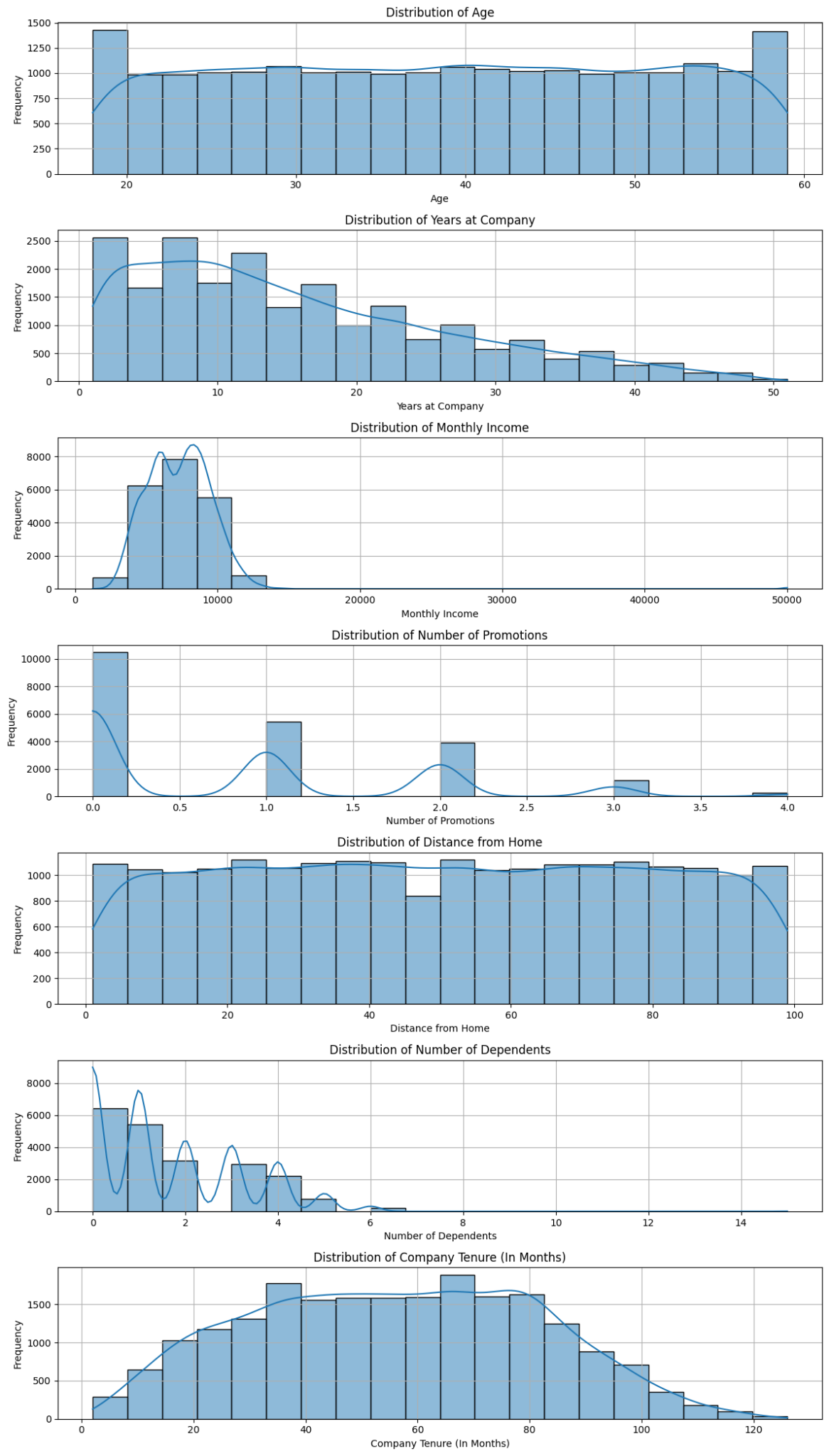


We see that the data is fairly balanced with a good proportion of data points in both the classes.

1. **Categorical Column vs Attrition:** Frequency of data from categorical columns was plotted against values of Attrition column.

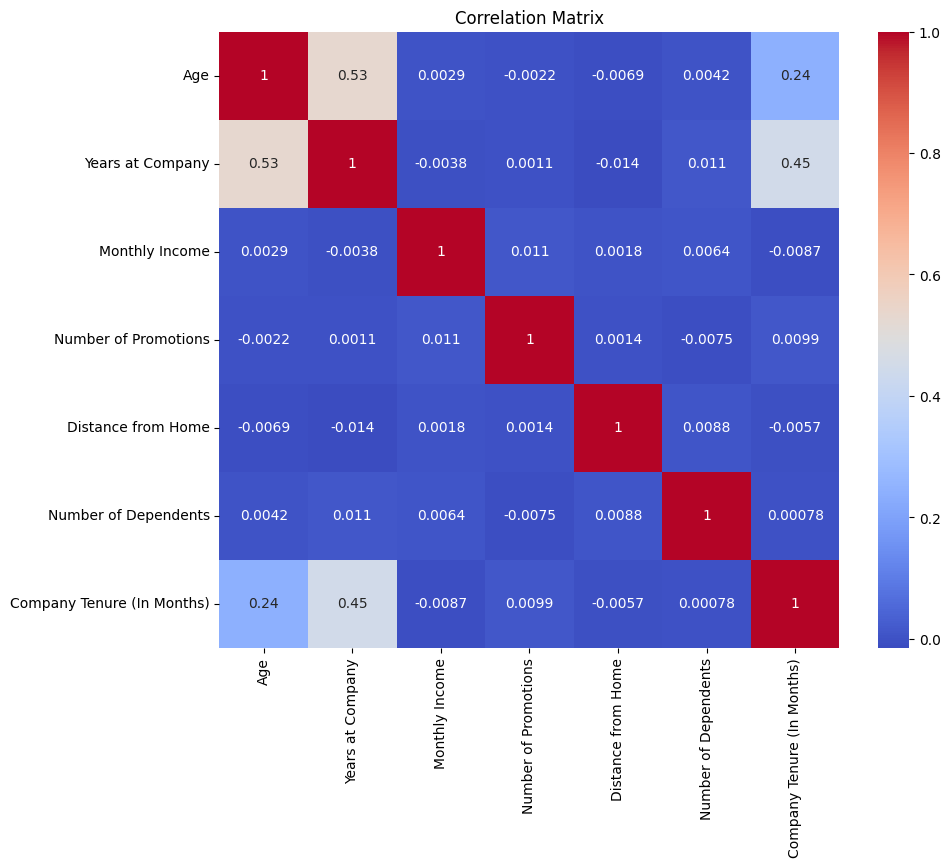


1. **Validate/differentiate EDA on train with EDA on validation [Optional]**
   1. **Univariate Analysis:** Distribution of numerical columns were checked for the validation data. Following are the distributions of all the numerical columns:



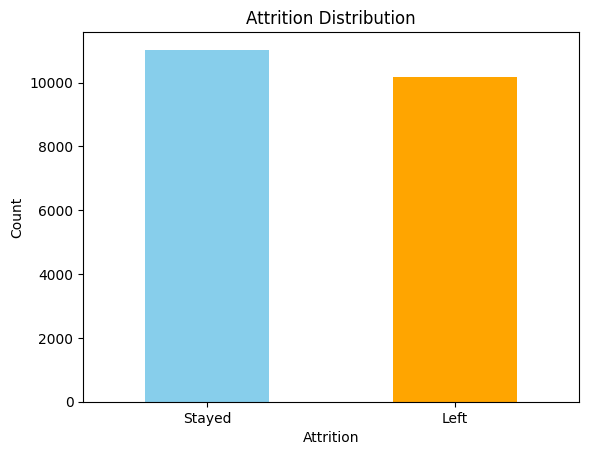
The distribution of numerical columns follows the same distribution as for training data.

* 1. **Collinearity:** Multicollinearity was checked among the numerical columns present in data.



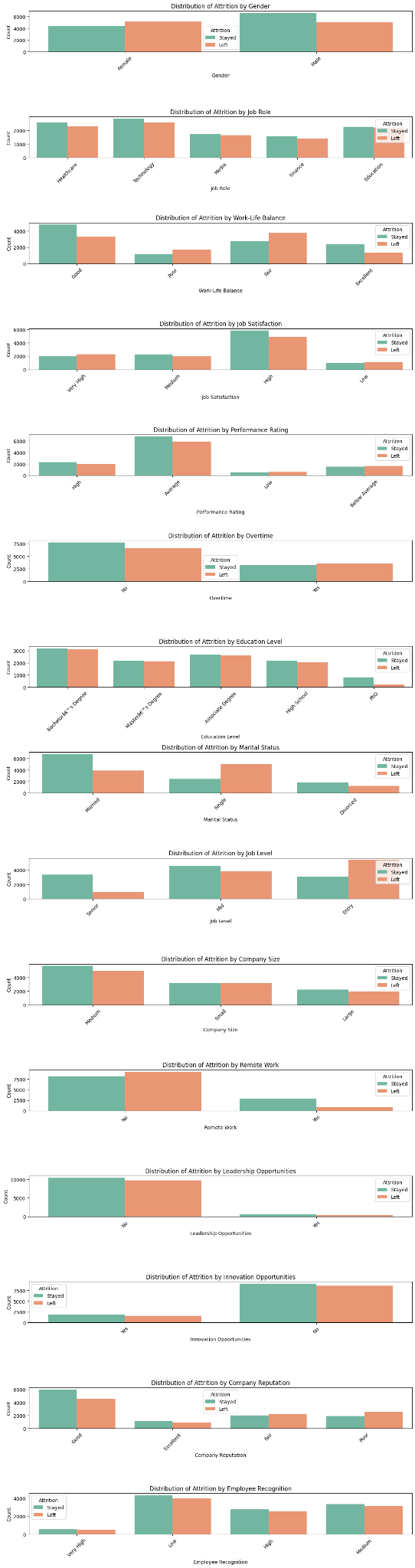
We see that similar to the training set, there is collinearity present between columns Age and Years at Company, and also between Age and Company Tenure.

* 1. **Data Balance:** Distribution for Attrition column was plotted in order to understand out of the total validation universe, how many people stayed and how many left the company.



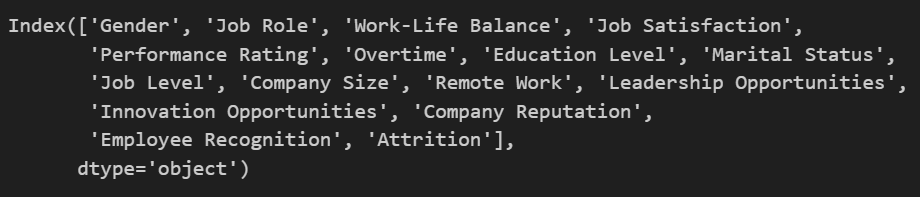
We see that similar to the training set, the data in validation is fairly balanced with a good proportion of data points in both the classes.

* 1. **Categorical Column vs Attrition:** Frequency of data from categorical columns was plotted against values of Attrition column.



1. **Feature Engineering**

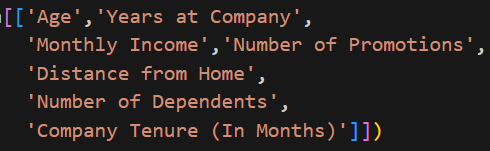
**Dummy Variable Creation:** For the categorical columns present in the dataset, dummy variables were created using the get\_dummies function. Dummy Variables were created for following columns:



After dummy variable creation, the original columns were dropped from the dataset.

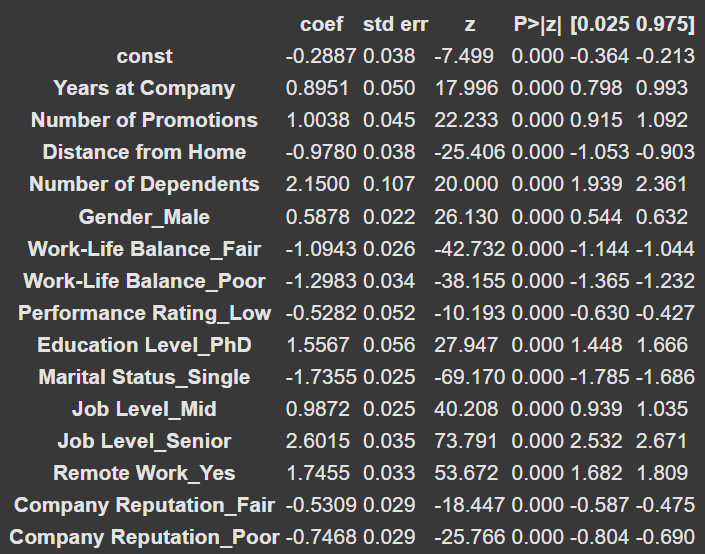
## **Model Building:**

* 1. **Feature Scaling:** Feature scaling was applied on the continuous/numerical columns of the dataset in order to bring the data within the same scale. Feature scaling was applied on the following columns using “MinMaxScaler”.



Correlations between different columns of the dataset were checked. There were no columns present which were highly correlated.

* 1. **Feature Selection:** After dummy variable creation, the dataset had over 42 features, which is a large number to deal with. In order to have a model built on the most important feature, Recursive Feature Elimination (RFE) was used to reduce the number of features to 15.
  2. **Building Model:** The logistic regression model was trained on these top 15 features. Following are the coefficients and p-values of the features basis trained model:

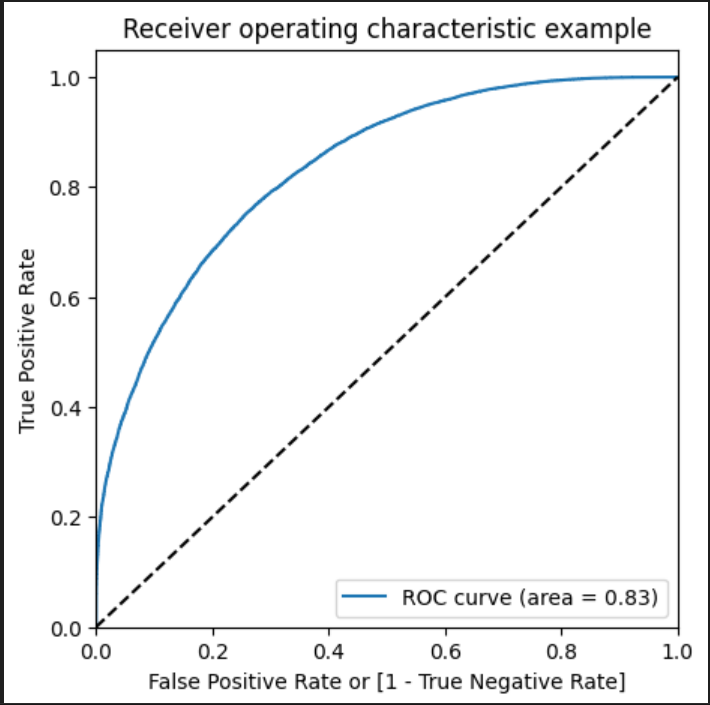


The table above provides us with the features used for building models along with coefficients of each of the features and their p-value. A positive coefficient indicates that an increase in the value of a feature would increase the odds of the event occurring, in our case an employee leaving the company. On the other hand, a negative coefficient means the opposite, i.e., an increase in the value of a feature would decrease the odds of the event occurring. The p-value in a logistic regression model is used to assess the statistical significance of each coefficient. Lesser the p-value, more significant the feature is in the model.

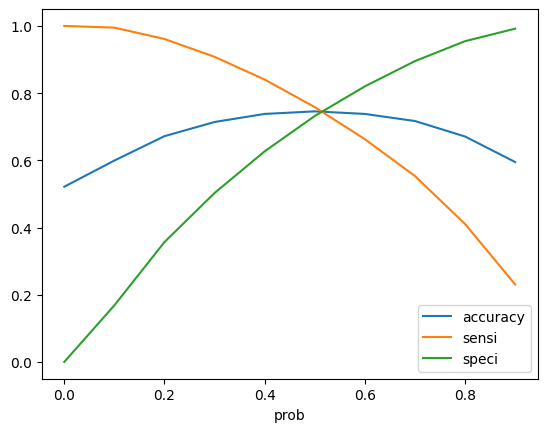


VIF was checked for the features in order to determine if multi-collinearity was present among the features. Since VIF value was less than 5 for the features, multi-collinearity is not significant in the features.

* 1. **Base Model Evaluation:** Predictions were made on the training set using the model built. The model had an accuracy of 74.57% on the training dataset with a sensitivity of 0.76, specificity of 0.73, precision of 0.75 and recall of 0.76 when the cutoff was set at 50%.
  2. **Finding Optimal Cut-off:** Model was fine-tuned by finding the optimal cutoff point. For this, first the ROC curve was checked, where the area under the curve was 0.83, which indicated we had a good model.

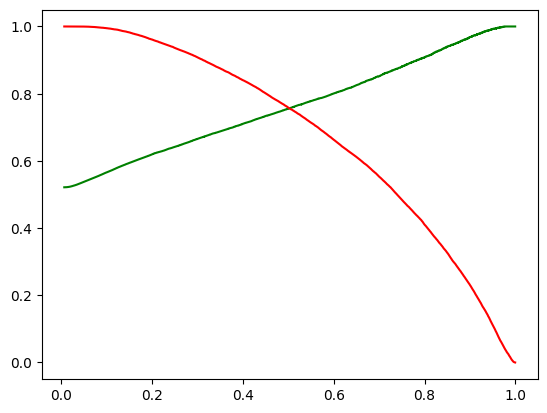


Next, a plot of accuracy, specificity and sensitivity of model for different cutoff values was created.



Basis the plot, we can see that the optimal cutoff is around 0.52.

Next, precision and recall of the model was checked. Precision of model is 0.76 and recall is 0.74. The precision recall curve was plotted to understand the precision recall tradeoff and determine the optimal cutoff point. Basis this approach as well, we get 0.5 to be the optimal cutoff.



Next, predictions were made basis 0.52 cutoff. Accuracy of model basis this was 74.63% with a sensitivity of 0.75 and specificity of 0.74. This model was further used for making predictions on the test set.

## **Prediction and Evaluation:**

The above model was used to make predictions on validation set. The model was able to predict with 74.3% accuracy on validation set with sensitivity of 0.74 and specificity of 0.75. The precision of the model on the validation set was 0.76 and recall was 0.74.

As the sensitivity and specificity of the model are relatively close, it indicates that the model is performing fairly well at distinguishing between positive and negative classes and there is a good balance in the model. Overall, the model is performing well over the test set, and can be used by the company.

# **Recommendations to Business**

The model will help the business to understand the likelihood of an employee leaving the organization. Basis this model, business can use following methods in order to reduce employee attrition:

* **Allowing Employees to Work Remotely:** In the model, Remote Work Yes feature has a coefficient of 1.7455. This means that employee retention is higher in a setting where remote work is available. Hence, having a remote work option would help in employee retention.
* **Improving Work Life Balance:** In the model, Fair and Poor work life balance has coefficients -1.0943 and -1.2983. It means that poor work life balance leads to employee attrition. Business should honor employees’ work life balance; this would lead to better employee retention.
* **Attention to Low Performing Employees:** Low performing employees have higher probability of leaving the company, which is evident through the coefficient of -0.5282 for Low Performance Rating feature in model. HR should reach out to such employees and understand the reason for low ratings, and how companies can help in the improvement of their rating in future.
* **Improving Company Reputation:** The coefficient for Fair and Poor company reputation from the model is -0.5309 and -0.7468, which indicated employees in Fair and Poor rated companies tend to leave the company. Measures should be taken in order to have and maintain a good rating for the company. Companies can enhance their facilities, work environment and work culture in order to help with rating along with ensuring their performance in the market.