**SALIFORT MOTORS EMPLOYEE CHURN**

**Problem Statement:**

Build a machine learning model to predict employee churn and the most important reason behind it so that the company can improve employee retention.

**Business Requirements:**

As a data specialist working for Salifort Motors, you have received the results of a recent employee survey. The senior leadership team has tasked you with analyzing the data to come up with ideas for how to increase employee retention. To help with this, they would like you to design a model that predicts whether an employee will leave the company based on their  department, number of projects, average monthly hours, and any other data points you deem helpful.

**Steps Followed:**

1. Important packages required for the project were imported.
2. The dataset was loaded and exploratory data analysis was done to get basic information about the dataset.
3. The columns in the dataset were checked for and the misspelled and other columns missing the snake\_case format were renamed so that all of them now followed a particular format.
4. Missing values in the dataset were checked for and found out that there were no such cases.
5. Next duplicate entries were checked for and 3008 such records were found.
6. Next the rows containing duplicate entries were checked for and the idea that those entries were legitimate was considered. However, the likelihood was found to be very less so the duplicated entries were dropped in the next step keeping only the first entry intact and saved into a new dataframe df1.
7. Now, a boxplot was plotted to identify any outliers in tenure. There were outliers detected in tenure from the boxplot.
8. Then, the no of rows containing outliers in tenure were detected. They were 824 in total.
9. Now, the no of employees who stayed back and left the company were found along with their respective percentages.
10. Next, plotted a stacked histogram to visualize the distribution of no of projects for those who stayed and those who left.
11. It was seen from the plot that anyone with 7 projects had left the company. This was confirmed manually as well with code.
12. Next, a scatterplot was plotted to visualize the relationship between average monthly hours and satisfaction levels to see how many left vs stayed back.
13. Another histogram was plotted to visualize tenure vs retention.
14. Now, a boxplot was plotted to see the relationship between tenure and satisfaction levels.
15. Two histograms were plotted to see salary ranges of individuals with short(<7) and long(>6) tenures and compared by keeping them in categories of low, medium and high.
16. Next, a scatterplot was plotted to see the monthly hours compared to promotions in last 5 years and how many of them left and stayed back.
17. Another scatterplot was plotted to visualize monthly hours by evaluation score and employee retention was checked.
18. A histogram was plotted to visualize the no of employees who stayed/left by department.
19. Lastly, a heatmap was plotted to see strong correlations between variables in the data.
20. Now, after finding the variables strongly correlated and how they each contribute to churn, it was time for us to select our model. The given task is a classification as we need to predict if an employee will leave or stay(more accurately binary classification). We can use either a Logistic Regression or a Tree Based model for our given task. I chose tree based model to see how it performs first.
21. For modelling purposes, a copy of the dataframe was made and the categorical columns were encoded to convert to numerical/ordinal type.
22. The dataframe was divided into X(features) and y(target variable).
23. Next, the data was split into train, test with a test size of 25%.
24. Then, a dictionary of hyperparameters was created and Grid Search was initialized to find out the best parameter values.
25. Now, the decision tree model was fit to the training data.
26. The best hyperparameter values were found out.
27. Next, the AUC score was calculated which was a strong one, which shows that this model can predict employees who will leave very well.
28. Next, I wrote a function that will help us extract all the scores from the grid search.
29. Next up, we used the function to get all the scores from the Grid Search. All of these scores from the decision tree model are strong indicators of good model performance.
30. Now, decision trees can be vulnerable to overfitting, and random forests avoid overfitting by incorporating multiple trees to make predictions. So I constructed a random forest model next.
31. The random forest model was fit to the training data.
32. The AUC score was calculated.
33. The best parameters were checked.
34. The scores for random forest model were collected and compared to those of the tree. The scores of the random forest were found to be better.
35. Defined a function that gets all the scores from a model's predictions.
36. Next, got the predictions on the test data.
37. Then, a confusion matrix was plotted to check the model’s performance through visualization.

**Key Findings and Recommendations:**

The key findings are:

1. There are two groups of employees who left the company: (A) those who worked considerably less than their peers with the same number of projects, and (B) those who worked much more. Of those in group A, it's possible that they were fired. It's also possible that this group includes employees who had already given their notice and were assigned fewer hours because they were already on their way out the door. For those in group B, it's reasonable to infer that they probably quit. The folks in group B likely contributed a lot to the projects they worked in; they might have been the largest contributors to their projects.
2. Everyone with seven projects left the company.

* If we consider satisfaction level and tenure, Employees who left fall into two general categories: dissatisfied employees with shorter tenures and very satisfied employees with medium-length tenures.
* Four-year employees who left seem to have an unusually low satisfaction level. It's worth investigating changes to company policy that might have affected people specifically at the four-year mark, if possible.
* The longest-tenured employees didn't leave. Their satisfaction levels aligned with those of newer employees who stayed.
* Two groups of employees who left based on evaluation score and monthly hours worked: overworked employees who performed very well and employees who worked slightly less than normal.

1. To retain employees, the following recommendations could be presented to the stakeholders: Most employees leave the company due to over work. To address this issue:

* Cap the number of projects that employees can work on.
* Consider promoting employees who have been with the company for atleast four years, or conduct further investigation about why four-year tenured employees are so dissatisfied.
* Either reward employees for working longer hours, or don't require them to do so.
* If employees aren't familiar with the company's overtime pay policies, inform them about this. If the expectations around workload and time off aren't explicit, make them clear.
* Hold company-wide and within-team discussions to understand and address the company work culture, across the board and in specific contexts.
* Consider a proportionate scale for rewarding employees who contribute more/put in more effort.