**Problem Statement**

You are working as a data scientist with HR Department of a large insurance company focused on sales team attrition. Insurance sales teams help insurance companies generate new business by contacting potential customers and selling one or more types of insurance. The department generally sees high attrition and thus staffing becomes a crucial aspect.

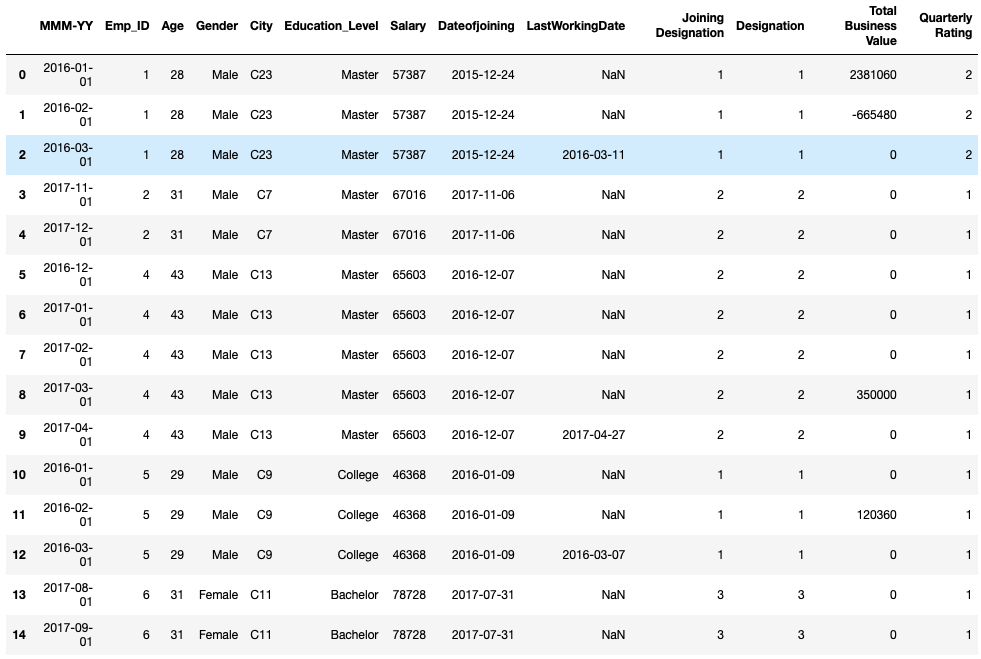
To aid staffing, you are provided with the monthly information for a segment of employees for 2016 and 2017 and tasked to predict whether a current employee will be leaving the organization in the upcoming two quarters (01 Jan 2018 - 01 July 2018) or not, given:

1. Demographics of the employee (city, age, gender etc.)
2. Tenure information (joining date, Last Date)
3. Historical data regarding the performance of the employee (Quarterly rating, Monthly business acquired, designation, salary)

Private Leaderboard Rank - 82  
Public Leaderboard Rank - 224

**Approach**

The main challenge was to transform the data provided into a Data Science Problem Statement. The snapshot of the original data is as below.



Objective was to predict if an employee will leave the organization in the upcoming two quarters. The target variable was taken such that if an employee leaves the organization then value would be considered as 1 and 0 otherwise.

**Feature Engineering**

The original 13 columns in the dataset are engineered/transformed to 16 features columns. The basic idea behind the feature engineering was that the following factors affected the churn of an employee:

1. Label encoding features like Education Level, Gender City
2. Calculating Tenure of the employee with organization basis Date of Joining and Last Working Date
3. Segregating Total Business Value into Business Earn and Business Loss
4. Change in Quarterly Rating and Salary over the tenure
5. Last Working Date feature transformed into Target variable

**Model Building**

AdaBoost algorithm was used along with 10 fold stratified sampling. The evaluation metric used was F1 score.

**Model Selection**

Along with AdaBoost, I did use models like Logistic Regression, Naïve Bayesian Model, Random Forest Model, Gradient Boosting Model. Even though F1 score was the metric used for evaluation, Precision and Recall parameter on Confusion Matrix were also referred while choosing the model for final evaluation. In training dataset the proportion of data pertaining to employee already resigned was less. Hence it would be a good reason to give importance to the prediction where employees who would be actually resigning should be predicted likewise.