**Predicting Employee Attrition**

**Solution Approach**

1. There was no ‘Target’ variable given so used the ‘LastWorkingDate’ column and first we replaced the null values by ‘0’ and then created a new variable "Attrition" and use a for loop to fill this variable by 0 & 1 indicating employee left or not respectively. Finally we merged that with our original data. Look at the code below:
2. Then we created a new variable ‘Experience’ which is the difference between ‘LastWorkingDate’ and the ‘Dateofjoining’. But before that we replaced the ‘0’ values in the ‘LastWorkingDate’ by current date which was ‘2021-11-20’ converted both therse variables into a datetime datatype and we get ‘Experience’ in days which we converted into years using basic python function.
3. We dropped the duplicates in the ‘Emp\_ID’ columns did a basic EDA on cleaned data.
4. Now we created a copy of our data and named it as ‘X\_train’ and by using the ‘Emp\_ID’ given in the ‘test data’ we removed those ID which were present in test data.
5. Similarly, we created another copy of our data and named it as ‘X\_test’ and we kept data of those ‘Emp\_ID’ which were present in ‘test data’. So now we have our train and test data.
6. Now we will create ‘y\_train’ and y\_test’ using only ‘Attrition’ column from ‘X\_train’ and ‘X\_test’ respectively and droppped the unnecessary columns like 'MMM-YY', 'Attrition', 'Emp\_ID' from our ‘X\_train’ and ‘X\_test’ data.
7. We converted ‘object’ datatype into numeric using LabelEncoder. Also we applied a StandarScalar to our data to improve the model performance.
8. Then we used ‘Logistic Regression’, ‘Random Forest Classifier’, ‘Naïve Bayes Classifier’, ‘KNN’ and ‘XGBoost’ and we used predictions of ‘XGBoost’ to submit the final submission where the initial score was ‘0.4316’ and after private leaderboard the score was ‘0.4316’.