**Approach For The Model**

**Data Cleaning and Feature Engineering:**

1. We see that there are multiple rows for the same employee with different Total Business Values hence we will group the data on the basis of Emp\_ID and add up the Total Business Value to get single record for Each employee.
2. For the ones who have already left where separated from the training dataframe.
3. A new feature for tenurity was made for all the employees, taking present day as 2017-12-31 as the present date .
4. One more new dataframe is made with all the details in it for every unique employee and adding the total business value for each of the employee.
5. Both the dataframe is then merged together.
6. A new feature is made from the both the dataframe to be considered as our Target variable named Attrition with values – 1( those who have left the organisation) and 0(for the ones who have not left it).
7. Columns **MMM-YY,Dateofjoining\_y , LastWorkingDate** as new features where already made with these and MMM-YY was not taking into consideration as important for building the model.
8. Converted the data type to datetime for **Dateofjoining\_x** to fill the na values in the column **Tenurity** of the remaining employees who are still in the organisation taking 2017-31-12 as the present working date.
9. Created new feature Promoted based on existing feature of Designation and joining Designation.
10. New Feature for **Joining day, Joining month, Joining year** is created from the Dateofjoinin feature in the dataset.
11. Changed the datatype Tenurity and removed unwanted string values from each of the value.

**Data-Visualization:**

1 . To check for **Outliers** in the dataset boxplot was plotted for every continuous feature and found out it is present in the features **Salary** and **Total Business Value** and **Age.** Outliers were then treated with z-score method.

2. Displot was plotted for continuous feature to check if there is any Skewness present in the dataset or not, again it was found on the features, **Salary, Total Business Value and Age which is practically not possible as well as in the Tenurity.** Skewness was then removed using quantile method from these feature.

3.Plotted heatmap inorder to see the co-linearity amongst the features. There was hardly any Co-relation found amoungs the features.

**Data- Preprocessing:**

1 - Dropped off the column Emp\_ID as after checking everyting are not found to be important.

2 - Removed the skewness from the features **Age, Salary, Total Business Value and Tenurity** with the help of quantile method.

3 – Removed outliers from **Total Business Value** feature with the help of z-score method.

4 – With the help of Label Encoder we performed label encoding on the feature Education\_Level as it Ordinal.

5 - With the help of get dummies method we performed encoding for the fetures Gender and City as these are nominal in nature.

6 - Separated label which is out Attrition in the dataset and features for model building into **X** for features and **Y** for label (**Attrition**)**.**

7 - Scaled the features for building our model with the help of Standard Scaler.

**Model Building:**

1 – Import all the algorithm and metrics required to build classification model and evaluate them on the basis of various metrics to find out the best model.

2 - Splitted the train set and test set into 4:1 ratio using train\_test\_split method.

3 - a ) **Random Forest** gave accuracy of 93.61% Accuracy, Precision of 94 %, recall of 94 % and F1 score of 94 %, Roc\_auc\_Score = 93.67%

b) **XGBClassifier** gave accuracy of 95.74% Accuracy, Precisio n of 96%, recall of 96 and F1 score of 96%, Roc\_auc\_Score = 95.78%

c) **CatBoostClassifier** gave accuracy of 95.41% Accuracy, Precision of 96%, recall of 95% and F1 score of 95%, Roc\_auc\_Score = 95.46%

d) **DecisionTreeClassifier** gave accuracy of 91.65% Accuracy, Precision of 92%, recall of 92% and F1 score of 92%, Roc\_auc\_Score = 91.66%

e) **LogisticRegression** gave accuracy of 96.23% Accuracy, Precision of 96%, recall of 96% and F1 score of 96%, Roc\_auc\_Score = 96.27%

3 - After evaluating all the models and plotting the roc\_auc\_curve it is clear that Logistic Regression is giving us the best results amongst all the algorithms that we have used.

4 - Checked the cross validation cross for each of the 3 model giving highest scores and we found out that Xtreme Gradient Boosting was overfitting the least among all of them.

5-We after the evaluation did Hyperparameter tuning using GridSearchCv on Xtreme Gradient Boosting.

After tuning XGboost classifier, unfortunately couldn’t get better results than the base model.

Hence we chose the base model of XGB and we used it to predict the **Target** of test data.