**Problem Statement**

Given the :-

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)

We need to find if the employee will leave the organization in next 6 months.

**Approach**

We will be training the model to capture the information about the employee’s attrition 6 months prior to the last working date. Below approach in the feature engineering, test set creation and modeling has been used:-

1. **Feature Engineer**

Since there is no target variable for our classification. We need to engineer the target variable from the the LastWorkingDate given for the employee. And also feature new variables from the given data. Below data processing has been done to create the final list of features:-

1. **Vintage**

It is the amount of time (in days) that the employee has spent in the organization. It was calculated with formula:-

Vintage = MMM-YY(reporting date) – DateOfJoining

1. **Promotions**

It is the number of promotions that user has received from the date of joining. It is calculated with formula:-

Promotions = Designation – Joining Designation

1. **Cumulative Total Business Value per Quarter**

Since negative value indicate refunds it must have been sold before to have been given a refund, therefore we have to calculate the cumulative total Business value per quarter. It might be directly related to the rating received by the employee in the next quarter.

1. **Happiness score**

Mostly when an employee leave the organization, it is because of the satisfaction with the salary being received by the user as compared to the people of the same age group with same level of education.

Formula to calculate happiness score:-

Mean = mean salary grouped by Age and Education Level

StandardDeviation = StandardDeviation in salary grouped by Age and Education Level

Happiness score(zscore) = (salary-Mean)/StandardDeviation

Higher the zscore more the happiness score. Negative value indicate employee is being paid lesser than the employee in the same age group and education level

1. **Motivation Score**

More motivated the employee more will be the business value that employee give to the organization. We will find the zscore of the cumulative business value per quarter grouped by the designation and quarterly rating.

Mean=Mean of Cumulative business value grouped by Designation and quarterly rating

StandardDeviation= StandardDeviation of Cumulative business value grouped by Designation and quarterly rating

Motivation score = (cumulative business value-Mean)/StandardDeviation

1. **Target**

Target value 1 is calculated by marking last 6 reported employee’s data as 1. Because we actually want to look for signs about employee leaving in the provided data 6 months prior to the last working date month.

1. **Train Test Split Data**

All the employees who have not left the organization, last 6 months records needs to be moved to the test file. Since we are not sure if the employee will be leaving the organization in the next six months. We are not sure that the Target for those rows will be 1 or 0. They are actually the test data whose predictions will be used for the final submission.

1. **Modelling**

Data was trained on 3 models:-

1. Random Forest
2. CatBoost
3. LightGBM

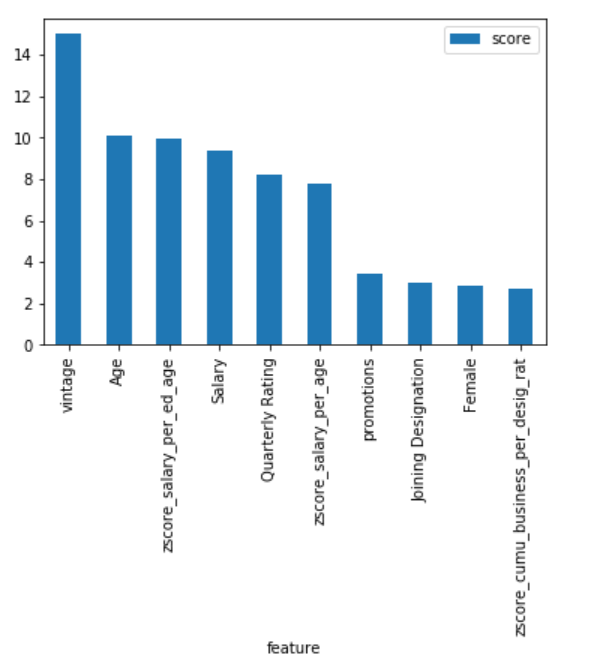
Their mean of prediction probability has been used to reach the final probability of Employee attrition.

Now since we have 6 records for each employee in the test data giving us the probability that the employee will leave the organization or not. We have calculated the mean of probability of last 6 months prediction.

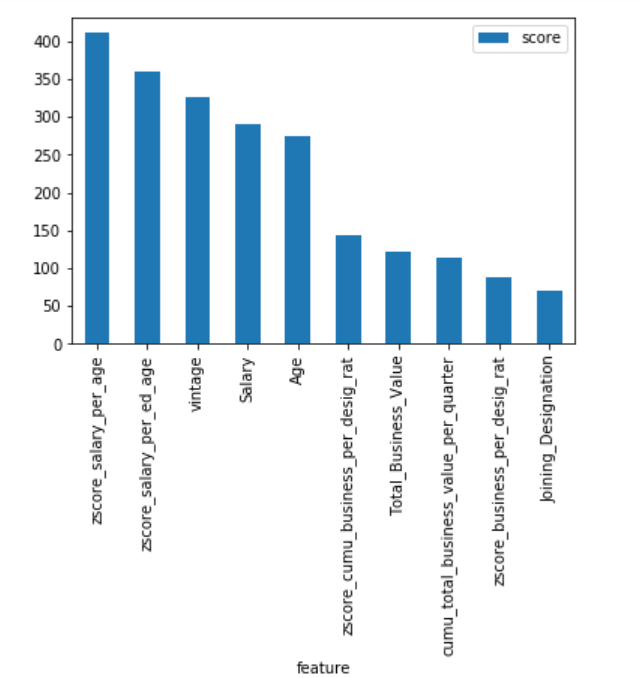
After tuning the threshold for the probability finally we got the best result at 0.43. Therefore any probability >= 0.43 will be given the target as 1(saying the employee will leave the organization within next 6 months).

**Feature Importance**

1. Below is the graph showing the most important features in catboost model creation

****

1. Below is the graph showing the most important features in lightgbm model creation



As can be seen the new features engineered are providing a great insight in the data and helping the model to reach a good level of accuracy.