Data preprocessing:

* There were a lot of discrepancies in the data like the scores which should have been between 0-1 were greater than 1 and given as percentage. So, these were preprocessed by bringing them within 0-1.
* It was observed that the missing values were created during 2013-105 and the data was complete for 2016-17
* Some important features like qualscore was imputed :
  + It was observed that qualscore was missing only for some performance years of some ACO and were highly similar for a ACO
  + Hence, it the data was sorted by the ACO\_name and the qualscore was imputed by backfilling by the last observed data.

The values where per\_capita\_expenditure was missing, those were dropped

Correlation and plot\_importance study:-

To understand the importance of features for per\_capita\_expenditure, two kinds of study was done:

Pearson Correlation: The ‘correlation.csv’ is the file having correlation values between target feature with other features

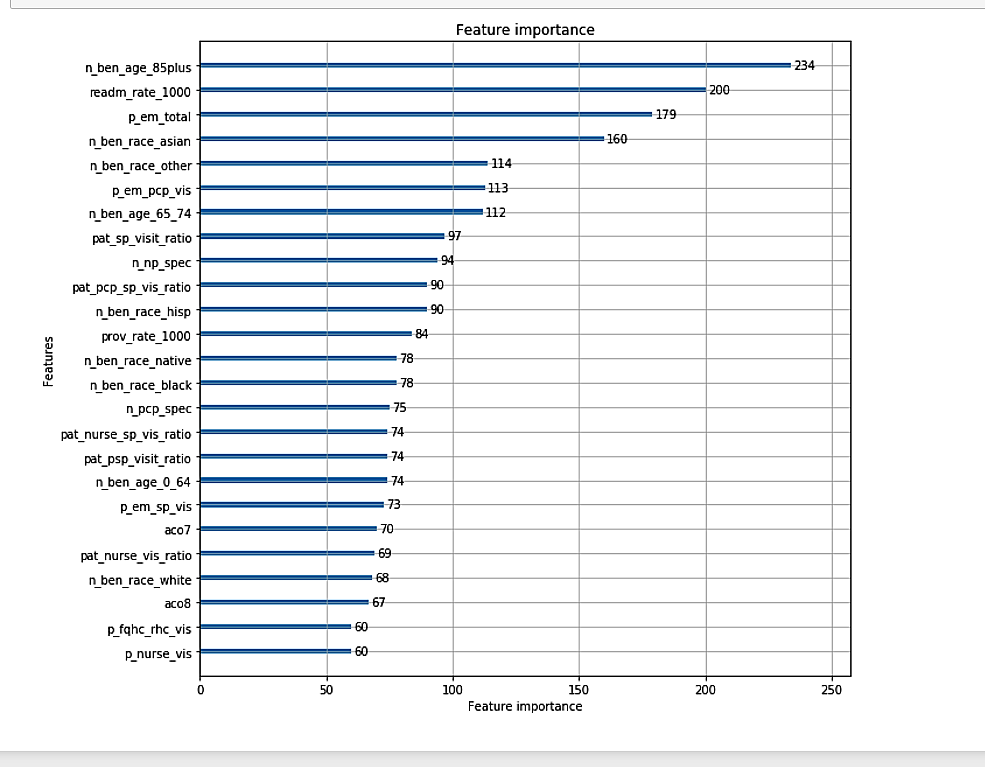
Plot\_importance : ‘DART’ (Dropout boosting tree) LightGBM was used to study the important features.

**R2\_score with crude data was :0.65**

**R2\_score after data engineering: 0.69**

It shows that now that newly built features were showing the variance of the data better and hence the features were reliable

Data Engineering:

1. pat\_psp\_visit\_ratio = train['p\_em\_pcp\_vis']/train['p\_em\_total'] : Ratio of primary healthcare visits
2. pat\_sp\_visit\_ratio = train['p\_em\_sp\_vis']/train['p\_em\_total']: ratio of specialized heathcare visits
3. pat\_nurse\_vis\_ratio= train['p\_nurse\_vis']/train['p\_em\_total']: ratio of visits to clinical nurses
4. pat\_pcp\_sp\_vis\_ratio= train['p\_em\_pcp\_vis']/train['p\_em\_sp\_vis']: ratio of visits to primary health care and specailised expert
5. pat\_nurse\_sp\_vis\_ratio= train['p\_nurse\_vis']/train['p\_em\_sp\_vis'] : ratio of visits to nurses and specialed experts
6. n\_np\_spec= train['n\_np']/train['n\_spec']: ratio of number of nurses and specilaied nurses
7. n\_pcp\_spec = train['n\_pcp']/train['n\_spec']: ratio of primary clinical nurses and specialized nurses
8. n\_np\_cns = train['n\_np']/train['n\_cns']: ratio of nurses and clinically trained highly trained nurses
9. 

Plot importance showing new features doing quite well than their constituents