## Student DataFest 2018: Data Identity challenge

* In this competition we have records from multiple trainees, with multiple records from each trainee. Same trainee ids are present in train and test dataset.
* So, we have to make a model which can understand data about each trainee as well as it should generalize well for some new unseen trainees.
* Test dataset has 922 trainees which are not present in Train dataset and total of 1,347 records from test data which are unknown (whose trainee id is not present in train dataset). In Test dataset out of 31,349 only 1,347 records are unknown.
* There are 19,422 unique trainee ids.
* We need a model which can learn train data and while prediction it should be able to do prediction based on trainee id.

NaN value treatment

* age and trainee\_engagement\_rating columns has NaN values.
* Mean is age is used to fill NaN values for age column based on each trainee\_id.
* Mean of trainee\_engagement\_rating is used to fill NaN values based on trainee\_id and program\_id.

Model

* I have used CatBoost for final model which works better than other boosting algorithms when we use categorical variable for training.
* Catboost doesn’t need extensive parameter tuning compared to other boosting algorithms and gives good results and chances of over fitting are low.
* In catboost there is no need to transform categorical variables to one hot encoded or label encoded variable because catboost handles them on its own. That’s why all the features are used as it is without doing label encoding or one hot encoding.
* Here I have also used trainee\_id and test\_id as a categorical features.
* All the features except ‘age’ and ‘total\_programs\_enrolled‘ are not used as a categorical variables.
* For catboost model all the default hyper parameters are used except eval\_metrics=‘AUC’ .
* Feature importance of model: 