**The approach towards the problem is as follows**:

1. Import the required libraries and data sets (train and test set)
2. Created the variables ‘train\_null’ and ‘test\_null’ to see the number of nans in each column of train set and test set respectively.
3. Defined two functions ‘my\_ohe’ and ‘my\_combine’.
   1. ‘my\_ohe’ is used to onehotencode a column and concat it to train\_set\_ohe and test\_set\_ohe corresponding to train set and test set respectively.
   2. ‘my\_combine’ is used to concat a column as it is to train\_set\_ohe and test\_set\_ohe corresponding to train set and test set respectively.
4. Now we handle each column one by one.
   1. ‘id’ – drop this column as no useful information for the classification model
   2. 'program\_id'- Drop column due to multicollinearity risk from ‘program\_type’
   3. 'program\_type' – use ‘my\_combine’ for it
   4. 'program\_duration' - use ‘my\_combine’ for it
   5. 'test\_id' - use ‘my\_combine’ for it [note that we didn’t drop this column as many observations are same in both train and test set. So, we can take advantage of data leak]
   6. 'test\_type' - use ‘my\_ohe’ for it
   7. 'difficulty\_level' – Create a dictionary ‘diff\_map’ based on the effect of each category in this column on the target value.
   8. 'trainee\_id' - use ‘my\_combine’ for it [note that we didn’t drop this column as many observations are same in both train and test set. So, we can take advantage of data leak]
   9. 'gender' - use ‘my\_ohe’ for it
   10. 'education' - Create a dictionary ‘edu\_map’ based on the effect of each category in this column on the target value.
   11. 'city\_tier' - Create a dictionary ‘city\_map’ based on the effect of each category in this column on the target value.
   12. 'age' – first fill the missing values using the median value in train set and use ‘my\_combine’.
   13. 'total\_programs\_enrolled' - use ‘my\_combine’ for it
   14. 'is\_handicapped' - use ‘my\_ohe’ for it
   15. 'trainee\_engagement\_rating' - first fill the missing values using the mode value in train set. Create a dictionary ‘engage\_dict’ based on the effect of each category in this column on the target value.
5. Convert train\_set\_ohe and test\_set\_ohe to matrices of independent vars (X\_train and X\_test respectively)
6. Let y\_train be the vector of target variable
7. Fit the classication model (catboost) as it is very useful for large number of categorical variables in the data set. Also, specify the indices that we need to onehotencode and use parameter tuning to attain the best possible CV score.
8. Predict the results and write to the submission file