Approach

I just followed a simple approach in which I **invested my 70-80% time doing feature engineering, data exploration and finding any hidden trends** if there are any in the dataset and **rest of the time is invested in model selection** and fine tuning the model.

1. First of all as stated above I just explored the dataset, in which I specifically looked for columns with null values.
2. Then I just plotted the output / target column to look for class imbalance which was not present in this case.
3. Then I plotted all the features individually, I was trying to observe that if there are any extra categories or trend present in training set that are not their in test data.
4. Then I created 3 features which are program_id_no, age_miss, educated
5. Then I **filled the NULL values in the age column with median() value** and **NULL values in the trainee_engagement_rating with -1.0**.
6. I tried different ML models like KNN, RF, XGBOOST but they didn't give me the results I was expecting so after that I tried CATBOOST and tuned it with different values for the parameter and it just improved my previous score.

Quality checks performed / Errors found:

No errors were found.

Feature Extraction

1. **Program_id_no extracted from Program_id**
2. **Age_miss: 1 for age with NULL value or else 0**
3. **Educated: 1 for no qualification and else 0**

Model choice explanation:

I tried different ML models like KNN, RF, XGBOOST but they didn't give me the results I was expecting so after that I tried **CATBOOST** and tuned it with different values for the parameter and it just improved my previous score.