A brief on the approach

Since this was a binary classification problem, I used XGBoost as it gave better result than other classification algorithms.

Data-pre-processing / feature engineering

1. The first thing I did was to impute missing values. Since the missing values were from categorical variables only, I imputed the missing values using mode.

2. Dropped column ‘ID’ as it does not contributes/not needed for model making.

3. Added a new column ‘Age’ which was derived from columns ‘Lower\_Age’ and ‘Upper\_Age’.

Age = (Upper\_age + Lower\_age)/2

4. Converted categorical features to numerical using pandas.get\_dummies()

5. Separated target variable and independent variable.

6. Applied StratifiedKFold as cross-validation because it keeps the ratio of labels in each fold.

7. Created model using XGBClassifier.