**AmExpert Coda Lab 2021**

**Data Cleaning**

Dataset have certain categorical and numerical columns with missing values so the categorical columns were imputed with most frequent strategy whereas numerical columns were imputed with median values. After this the numerical columns were also treated for outliers by capping least value to 0.01 percentile and highest values to 0.99 percentile.

**Encoding**

So created different datasets with Label Encoder , One Hot Encoder , Cat Encoder and WOE(Weight of evidence encoders) on the categorical columns.

**Feature Enginnering**

Created bin features on the numerical columns along with log transformed features too.Created aggregated features of sum,max,min,mean of numerical columns. A have child column for whether the customer has a child or not also no. of adults in the family , a categorical column that if he has been a defaulter earlier on irrespective of the duration, ration of yearly debt payment with credit used and also along with yearly debt payments with net\_yearly income.

Also created a Kmeans features with 5 cluster to be labelled in the dataset.

**Models**

As the target value was skewed , I used Stratified K fold for the dataset, Majorly all the tree models performed greatly for the problem catboost classifier being the best one and then lgbm and XGBoost .

Created all the predictions from the models stated and then took the mode of all the prediction in a row for all the models i.e same as a voting classifier.

**Tools Used**

Anaconda, Jupyter Notebook ,Spyder