**Problem statement**: Predict whether a customer is lead or not (Interested or not) for credit card based on his/her details and their relationship with Bank.

**Approach:**

First step is to look at data provided by Bank (Both test and train) and see whether they have same distributions. There were missing values in Is Credit column and I have given another label to them as it is not appropriate to delete the rows or that particular column. Now the data do not have any missing details. I created another column “Is Credit null” with ones and zeros to identify the null values in “Is Credit” column. Next step was to make new features from ‘Region code’ as they cannot be directly used. I made new features by aggregating this region code feature with other columns. This has created many new feature columns and few of them may be highly correlated and redundant, so I have removed those features which are having high (>0.95) correlation with other. Since the Average account balance distribution is heavy tailed with right skew, I have transformed it with log function.

I was very careful not to create many features as it can affect the model Performance and also it is difficult to interpret the features at later stages.

Now the data preprocessing is done and ready for training.

**Training:**

I have tried many algorithms, But I found LGBM Classifier and CAT Boost classifier with cross validation has outperformed other models. I have shown Feature importance’s of each feature column based on the model and removed few features which are not contributing to the prediction.