1. **A brief on the approach, which you have used to solve the problem.**

I have proceeded with an ensembling technique wherein I have used the catboost algorithm along with light gbm and xgboost to get the probability of whether a client will be a successful lead or not. The output of each of the three algorithms are ensembled using appropriate weightage for each of the techniques to get the final predictions.

Raw Data

Data cleaning : Convert missing data(credit\_product) to a new category(not\_given)

New Feature: not\_given

Creation of interaction features for all categorical variables

Hyperopt: best\_parameters

Catboost

Light GBM

OOF: Ensemble 8

XGBoost Classifier

Blended model:

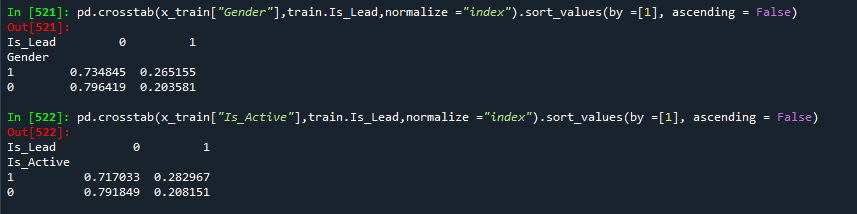
w(i)Catboost\_pred + w(j)LGBM\_pred + w(k)XGBM\_pred

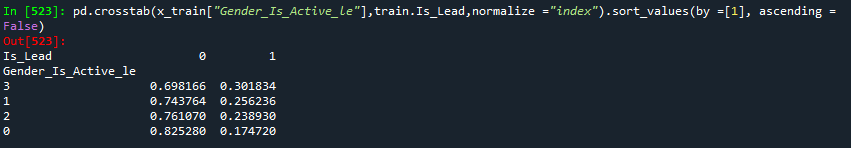
w’s are the weightage of individual models which are multiplied with predictions of each models

1. **What data-preprocessing / feature engineering ideas really worked? How did you discover them?**

There were missing values in Credit\_Product column . I have ran the models by replacing all the missing values with another class “not\_given” and without replacing the missing values as well. Surprisingly the score without replacing the missing values were better.

I have made interaction features by combining every two categorical features and performing frequency encoding as shown below in the screenshot, both the **Gender** and **Is**\_Active sorted records have around 73 % 26 % distribution with respect to the target Is\_**Lead**.





When the Gender and Is\_Active columns **interacted,** the distribution for target -1 is higher than the individual Gender and Is\_Active categorical columns in the new variable **Gender\_Is\_Active** . Creating interaction features for all the categorical columns based on the above analysis helped me to increase my Score from 0.803 to 0.8723 .

1. **What does your final model look like? How did you reach it?**

The final model is a blend / ensemble of the three models catboost, lgbm and xgboost with interaction features introduced amongst all the categorical variables in all the models. Each of the models predicts the probability of each of the customers . The probabilities for every customer of each model is given appropriate weightage and the final prediction is obtained as a weighted combination of the three models.