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

We have to predict whether a customer will be interested in the Credit Card. The given dataset has 7 Categorical Features and 3 Numerical Features excluding the target variable. We checked for missing values in the data and found 12% of missing values in the Credit\_Product column. Then tried to figure out the reason for the missing values by exploring the data and performed the missing value treatment. We dropped the “ID” column from the data because it has high variability.

Since we have to classify the final result so tried a few simple classification algorithms (Logistic regression, Ridge classification) but they didn’t perform well so I go with tree-based algorithms, random forest and gradient boosting algorithms were overfitting the data so I finally came up with Cat Boost algorithm to build the final model.

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

First of all, we removed the “ID” column as most values are unique for this feature. While exploring the Credit\_Product column I noticed there are some irregularities along with null values. but we can not say they are missing values because this feature having null values is obvious, it means we don't know if the customer has any active credit product and some customers don't hold any credit product. Both the cases are valid business cases we need not do explicit missing value treatment. We can just fill it with some other value to let the model know its pattern(missing value has a lot of info here). For the sake of EDA, I replaced all the missing values with 'Undefined'. Later in the pre-processing, I replaced 'Undefined' with some large negative value for both train and test data so that model identifies it easily.

We didn’t use any kind of encoding here since our model can perform well with categorical features.

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

I have used Cat Boost Algorithm in my final prediction model.

I started with trying simple models to understand the nature of data but didn’t get good results. so I moved on with tree-based models but they were over-fitting the data. Then I tried the cat boost algorithm and it performed well on the data, so I decided to go on with it and tried setting parameters to get the best result from the algorithm. This model returned a roc\_auc score of 0.8740734446188525.