I have used catboost classifier for my final model with 500 iteration, 0.02 learning rate, depth of the tree 12, Are under the curve to find accuracy, bagging temperature as 0.2, overfitting detector as Iter, od\_wait(no. of iterations to continue after optimal metric value)

**Data-Preprocessing:**

1. First on checking the data types of columns Gender, Region\_code, Occupation, Channel\_code, Credit\_Product, Is\_Active needs to be converted to category data type.
2. On checking the missing values, Credit\_Product alone has 12% of missing values
3. On checking the univariate analysis for numerical columns,

**Age:**

* Median age is 43
* Most of the customers are between age 29 to 58
* kurtosis = -0.44; very less likely to have extreme values
* skewness = 0.62; age is slightly biased to younger age

**Vintage:**

* Median vintage is 32
* Most customers are joined between 14 to 79 months
* kurtosis = -0.7; very likely to have extreme values
* skewness = 0.79; vintage is slightly biased to new customers.

**Avg Account balance:**

* Median vintage is 1128403.1
* Most customers have balance between 275466 to 1981339.
* kurtosis = 14.3; Extreme values are present.
* skewness = 2.97; significantly biased towards lower account balance.

1. Next univariate analysis for categorical columns,

**Gender:** Male customer accounts for 1.2 times more than female.

**Region\_code:** Out of the 35 regions, 5 regions alone account for 55% of customers.

**Occupation:**

* Majority of people are self\_employed.
* There are extremely few Entrepreneur. Might explain Outlier/Extreme values in credit/debit.

**Credict product:** Majority of the customers are currently not having any credit products

**Is active:** Majority of the customers are not active

**Is lead:**

* Most of the customers are not interested in the credit card.
* Customers who are not interested can be explained by less number of active customers.
* From the percentage values, the dataset is highly imbalanced.

1. Directions got from univariate analysis are
   1. customer\_id can be dropped.
   2. Occupation can explain the higher avg account balance customers.
   3. Lower avg account balance can be explained by the high number of inactive customers.
   4. Person with no credit product can be explained with the occuption and inactivity.
2. On checking bivariate analysis for categorical variables
   1. In Gender, there is no correlation with any gender as both the gender produces similar count
   2. With region\_code, some region(around 6) accounts for more customers.
   3. With occupation, entrepreneurs are more interested in credit cards than other occupation category.
   4. Channel code, doesn’t provide any new information.
   5. Customers with Credit\_Product are more interested in getting credit cards than customers without it.
   6. As a surprise, number of interested inactive customers and interested active customers are almost similar.
3. On checking with the numerical bivariate analysis, there are no correlation between the numerical columns.
4. On checking the outliers in average account balance, almost customers with all the occupation has outlier values in which we can see Other and Entrepreneurs have more outlier values. But the median Avg\_account balance is more for entrepreneurs. Also the credit card interested people, have fatter tails.
5. For lower account balance with inactive customers, both active and inactive customers have similar account balance plots i.e both the categories have same lower and almost same higher amount. Also the credit card interested people have fatter tails.
6. No credit\_product vs in active vs occupation,
   1. majority of non credit product holders are in self-employed, salaried and other categories.
   2. some of the inactive accounts also has credit products
7. Then have imputed the missing values in Credit\_Product with the group specific mode of occupation in train and test data.
8. From our analysis, can see Credit\_products, Is\_Active, Occupation, avg\_account\_balance have significant effect on the target variable.

**Model:**

1. Initially checked with random forest model with rfc metric as gini, number of estimators as 100, which resulted in 0.7541 roc-auc score. Important feature identified here are average account balance, vintage, age.
2. On trying with logistic regression, roc-auc score is 0.7007
3. Next tried with adaboost classifier, roc-auc score is 0.764. Here important features identified are age, vintage, channel code\_X3, channel\_code\_X2.
4. With catboost classifier, roc-auc score is 0.788 in validation set. Important feature identified are occupation\_salaried, age, vintage, credit\_product\_yes, is\_active\_yes
5. With xgboost, got the auc-roc as 0.779 and important feature are age, vintage, avg\_account\_balance, occupation\_salaried.

From this, based on roc-auc score, I chose, catboost classifier. Now can perform grid search to tune the hyperparameters.

On sending with the parameter grid as {'depth':[6,5,7,8,9,10,11,12,13],

'iterations':[250,100,500,1000],

'learning\_rate':[0.03,0.001,0.01,0.1,0.2,0.3],

'l2\_leaf\_reg':[3,1,5,10,100],

'border\_count':[32,5,10,20,50,100,200]

}

Fount the best parameter combination as