Technical Assessment

DANIEL SOUSA

22/07/2022

Data Understanding/Data Preparation

Interesting columns/Preprocessing

YEAR_MONTH – Values range from 202101 to 202106, not a representative sample of a year; It was processed to only appear the month digit.

Customer_Age – Two customers have ages: 235 and 218, data integrity In question, such records were removed;

Gender - If a news comes out that a bank is using Gender to predict which customers are more likely to churn or not, I am sure that it would be very damaging to the reputation of the bank;

Data Understanding/Data Preparation

Interesting columns/Preprocessing

Education_Level - About the education level I would say that it is a developed country;

Marital_Status - A very low percentage of divorces happen in the data;

Total_Product_count – 22% of missing entries. For the propose of the interview, I just filled the missing values with the mode. A better approach would be to use for instance the IterativeImputer from sklearn which models the feature with missing values as a function of other features, and uses that estimate for imputation. Also I will use LightGBM as one of the models, it can deal with missing features directly it will ignore missing values during a split, then allocate them to whichever side reduces the loss the most;

Data Understanding/Data Preparation

Interesting columns/Preprocessing

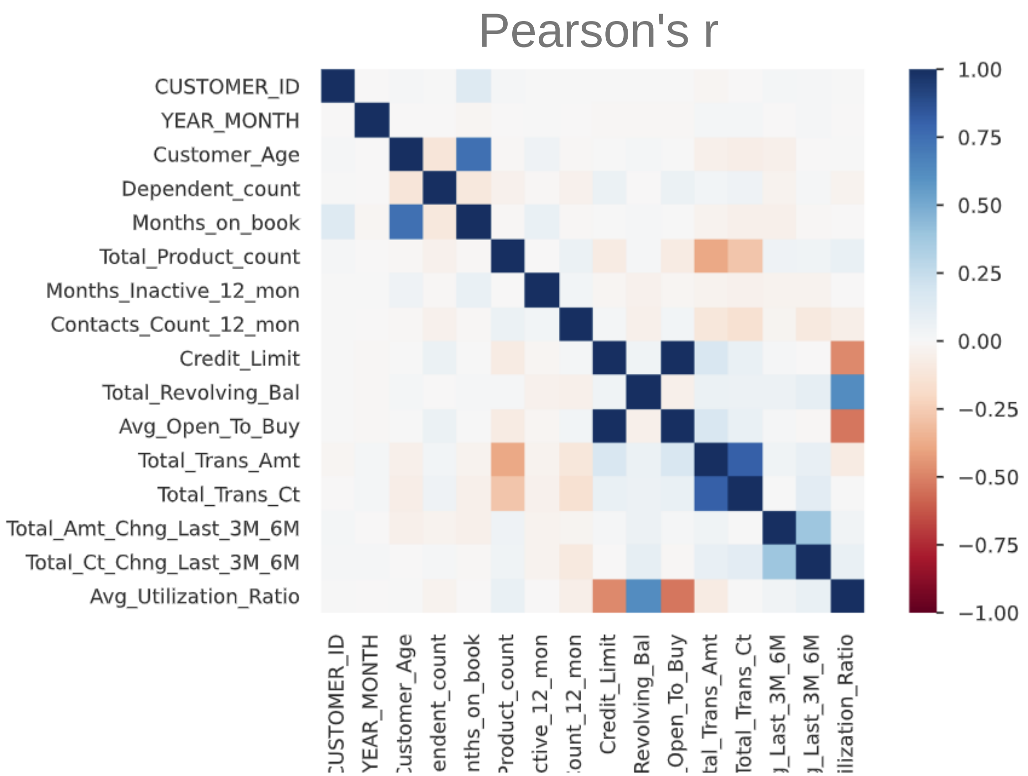
Total_Amt_Chng_Last_3M_6M, Total_Ct_Chng_Last_3M_6M – Are two highly correlated features, which makes sense as they represent two different ways of measuring the same quantities (changes between the last 3 months and the 3 months before);

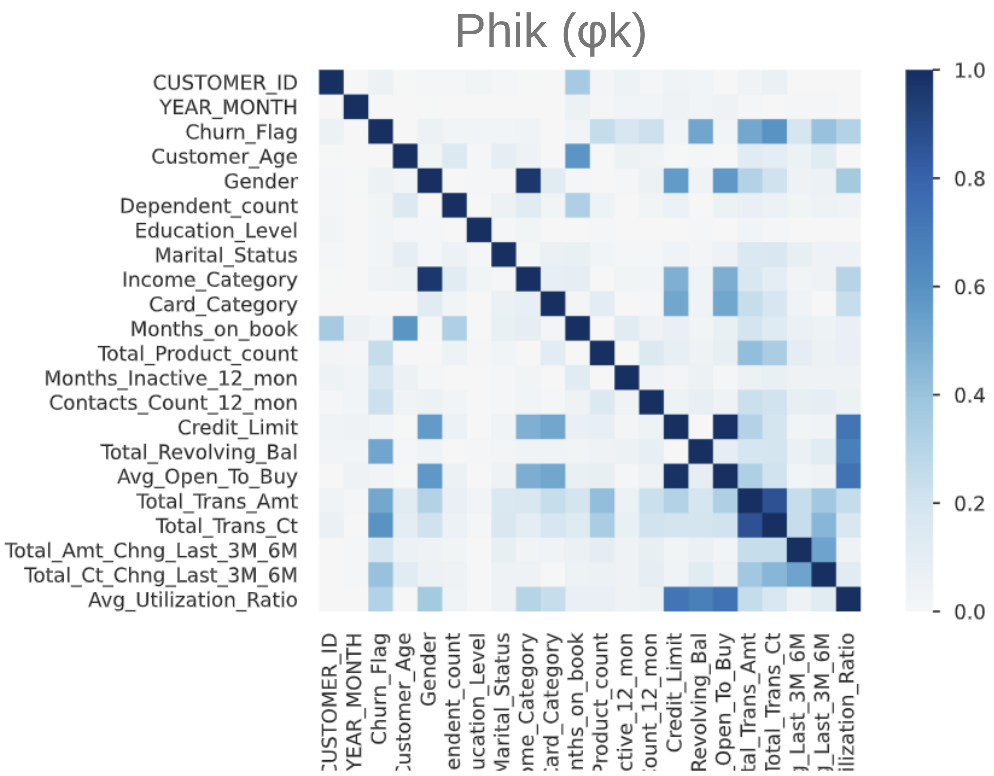
Total_Revolving_Bal, Total_Trans_Amt, Total_Trans_Ct - Customer churn is by definition when someone chooses to stop using your products or services. In effect, it’s when a customer ceases to be a customer. Therefore, it is only normal that the fields Total_Trans_Amt, Total_Trans_Ct are highly correlated with the churn label. However, do customers churn because of their Total_Trans_Amt, Total_Trans_Ct being low/high or they have a low/high Total_Revolving_Bal, Total_Trans_Amt, Total_Trans_Ct because they are no longer clients of the bank?

Data Understanding/Data Preparation

Target/Correlations

Churn – Highly correlated with three other features. Unbalanced, only 16% records are churned.





Dataset #1

Goal: Understand Feature importance/get baseline

Cyclical encoding

X

MONTH	Customer_Age	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Product_count
NUM	NUM	NUM	CAT	CAT	CAT	CAT	NUM	NUM

Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Trans_Amt	Total_Trans_Ct	Total_Amt_Chng_Last_3M_6M
NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM

Total_Ct_Chng_Last_3M_6M	Avg_Utilization_Ratio
NUM	NUM

Y

Churn_Flag
CAT

Test

Class	Count
0 (Existing Customer)	1276
1 (Attrited Customer)	244

Train+validation

Class	Count
0 (Existing Customer)	7224
1 (Attrited Customer)	1383

Dataset #2

Goal: Understand Feature importance/get baseline

Y

Churn_Flag
CAT

X

MONTH	Customer_Age	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Product_count
NUM	NUM	NUM	CAT	CAT	CAT	CAT	NUM	NUM

Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Trans_Amt	Total_Trans_Ct	Total_Amt_Chng_Last_3M_6M
NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM

Total_Ct_Chng_Last_3M_6M	Avg_Utilization_Ratio
NUM	NUM

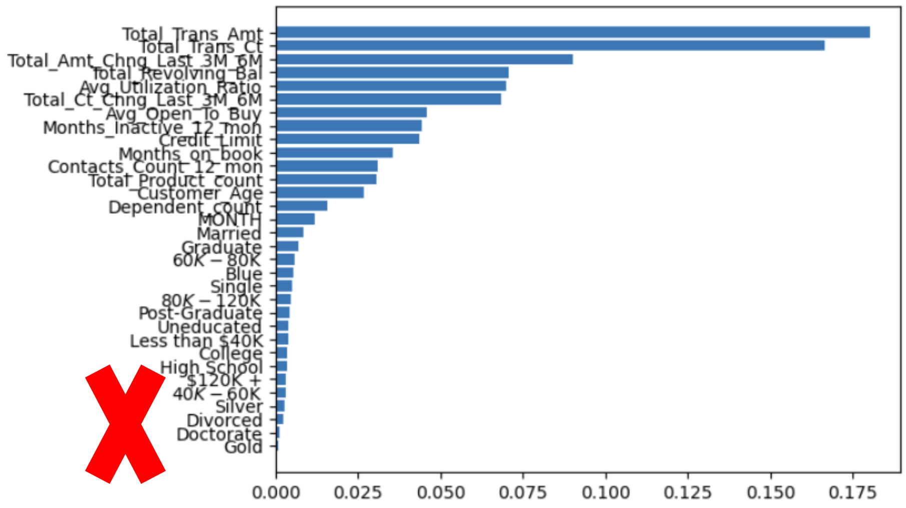
	Customer_Age	Months_on_book	Credit_Limit	Avg_Open_To_Buy	Total_Trans_Ct	Total_Trans_Amt
Customer_Age	1.000000	0.837697	0.080026	0.088829	0.199205	0.209906
Months_on_book	0.837697	1.000000	0.074322	0.072407	0.129879	0.149779
Credit_Limit	0.080026	0.074322	1.000000	0.987058	0.198485	0.323830
Avg_Open_To_Buy	0.088829	0.072407	0.987058	1.000000	0.205575	0.344228
Total_Trans_Ct	0.199205	0.129879	0.198485	0.205575	1.000000	0.856598
Total_Trans_Amt	0.209906	0.149779	0.323830	0.344228	0.856598	1.000000

Train+validation

Class	Count
0 (Existing Customer)	7224
1 (Attrited Customer)	1383

Test

Class	Count
0 (Existing Customer)	1276
1 (Attrited Customer)	244



novobanco

Tested models

Bagging Methods

BRF: under-samples each boostrap sample to balance it

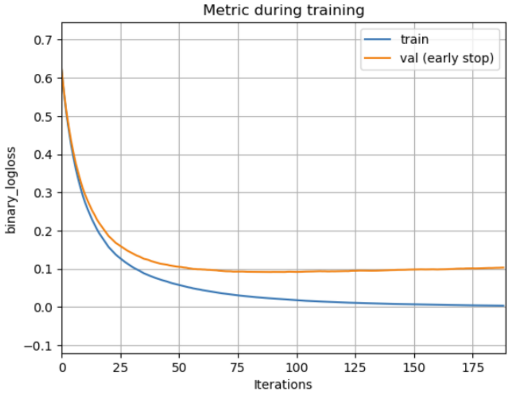
	Baseline (RF) Dataset V1	Model 1 (BRF) Dataset V1	Model 2 (BRF) Dataset V2
			
	$F_1 = 0.813$	$F_1 = 0.827$	$F_1 = 0.770$
Key hyperparameters:	n_trees=10	n_trees=10 cls_weights=balanced	n_trees=10 cls_weights=balanced
	#Features:32	# Features:32	#Features=11

Tested models
Boosting Methods

min_child_sample=It requires each leaf to have the at least the specified number of observations so that the model does not become too specific.
feature_fraction=If you set it to 0.8, LightGBM will select 80% of features before training each tree. (Randomly)

$$\text{Loss} = -\frac{1}{\text{output size}} \sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

Model 3 (LightGBM) Dataset V1

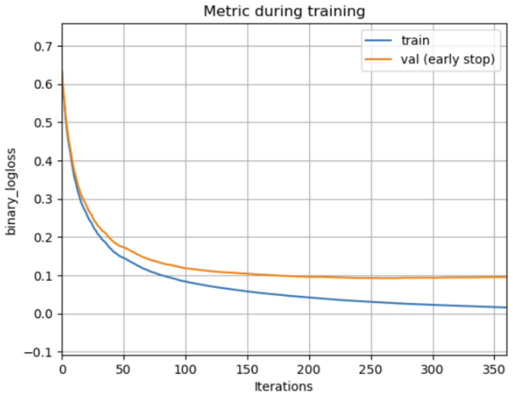


$F_1 = 0.934$

n_trees=89 (early stopping)
cls_weights=balanced

#Features:19 (4 categorical)

Model 4 (LightGBM) Dataset V1



$F_1 = 0.921$

n_trees=260 (early stopping)
min_child_sample=400
feature_fraction=0.8
cls_weights=balanced
#Features: 19 (4 categorical)

Key hyperparameters:

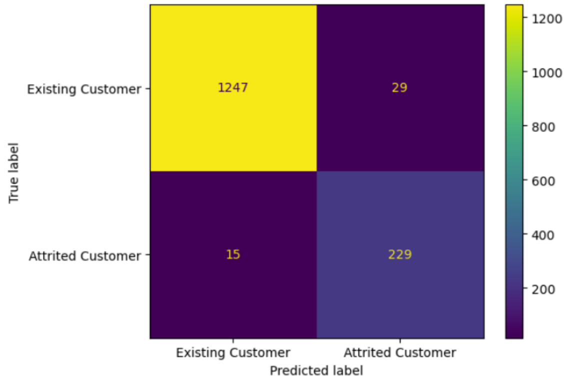
Tested models
Boosting Methods

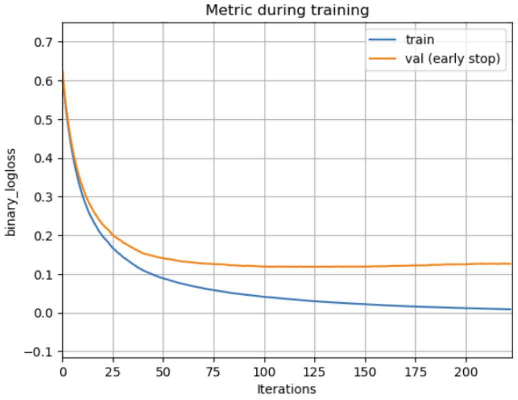
min_child_sample=It requires each leaf to have the at least the specified number of observations so that the model does not become too specific.
feature_fraction=If you set it to 0.8, LightGBM will select 80% of features before training each tree. (Randomly)

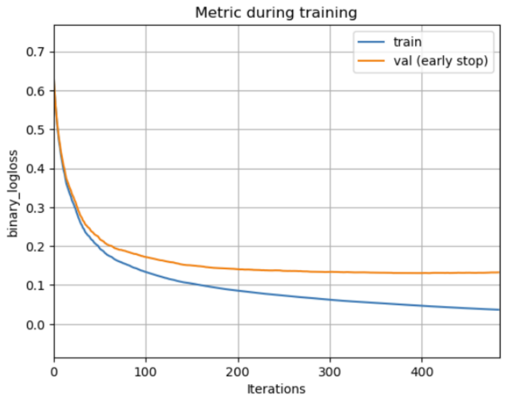
Model 5
(LightGBM)
Dataset V2

$$\text{Loss} = -\frac{1}{\substack{\text{output}\\ \text{size}}} \sum_{i=1}^{\substack{\text{output}\\ \text{size}}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

Model 6
(LightGBM)
Dataset V2







$F_1 = 0.912$

$F_1 = 0.870$

Key hyperparameters:

Model 5:
n_trees=123 (early stopping)
cls_weights=balanced

#Features: 10 (0 categorical)

Model 6:
n_trees=384 (early stopping)
min_child_sample=400
feature_fraction=0.8
cls_weights=balanced
#Features: 10 (0 categorical)

Future work

- Some further data engineering for creating some other features/reducing dimensionality, for instance, PCA;
- Improve the Total_Product_count column by creating a model to predict the value based on the other columns, right now if this went into production it would be important to understand if it is expected that in the future this value might keep coming like this (with NaN values);
- Some Data Upsampling techniques could be tested such as Synthetic Minority Oversampling Technique (SMOTE) could also be considered to increase the number of examples of the minority class.
- Organized hyperparameter search for instance using GridSearchCV or RandomizedSearchCV, would for sure improve model performance;
- Get input from the business on features that were considered;
- Regarding the LightGBM, the model is still overfitting slightly, some improvement should be executed on its hyperparameters tuning to improve this;
- Still regarding the LightGBM it's possible to use it with NaN values existing in a column it will ignore missing values during a split, then allocate them to whichever side reduces the loss the most;
- The evaluation should be done on an independent test dataset, right now the dataset that I used for testing was processed in the sense that there are examples that had NaN in the Total_Product_count. In the future one should get examples that had this column complete so that the test dataset does not suffer any kind of imputation;
- Other models could be tested like SVM and simple regressions. SVM tends to work better with smaller datasets but I don't think that we have a problem with that in our case.