Lazy FCA

National Research University "Higher School of Economics“

December 13, 2022

Student: Sakhnenko Vladislav

Dataset:

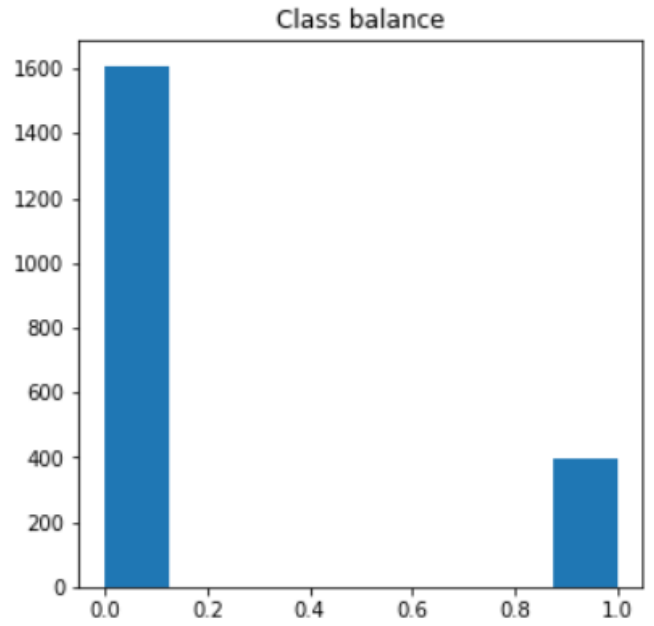
- Source: Kaggle.
- Description: Every bank wants to hold their customers for sustaining their business.
- Purpose: to predict the Customer Churn for ABC Bank

Size: 2000 examples

Metrics:

- F1 score
- ROC AUC

Accuracy is not suitable because the classes are not balanced



Example of a record in a table:

	credit_score	country	gender	age	balance	products_number	credit_card	active_member	estimated_salary	churn
0	596	Germany	Male	32	96709.07	2	0	0	41788.37	0

Binarizing features

1. Categorical features: country, gender, products_number, credit_card, active_member
2. Numeric features: credit_score, age, estimated_salary
3. Label: churn

- Nominal scaling is used to binarize categorical features
- Ordinal scaling is used to binarize numerical features. The numeric line of the attribute values is divided into intervals.

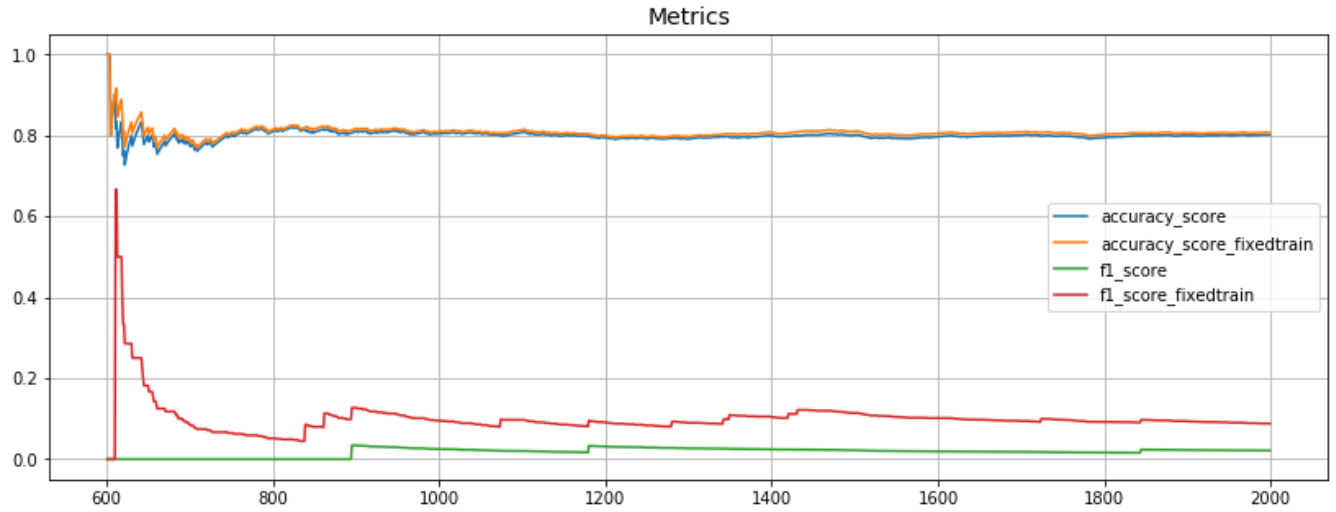
After binarization, the number of features increased from 9 to 33

	country: France	country: Germany	country: Spain	gender: Female	gender: Male	products_number: 1	products_number: 2	products_number: 3	products_number: 4	credit_card: 0	...	age_2	age_3
0	False	True	False	False	True	False	True	False	False	True	...	False	False

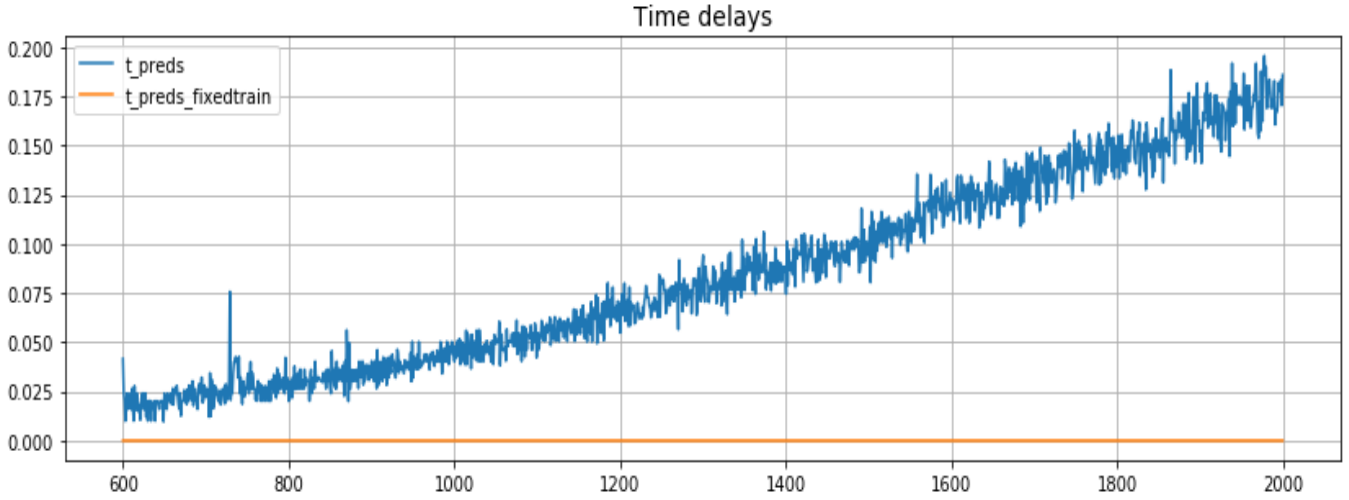
Baseline model

	precision	recall	f1-score	support
False	0.80	1.00	0.89	1119
True	1.00	0.01	0.02	281
accuracy			0.80	1400
macro avg	0.90	0.51	0.46	1400
weighted avg	0.84	0.80	0.72	1400

roc auc score: 0.5053



- The working time of the algorithms is about 2 minutes.
- Metrics are close to random prediction.



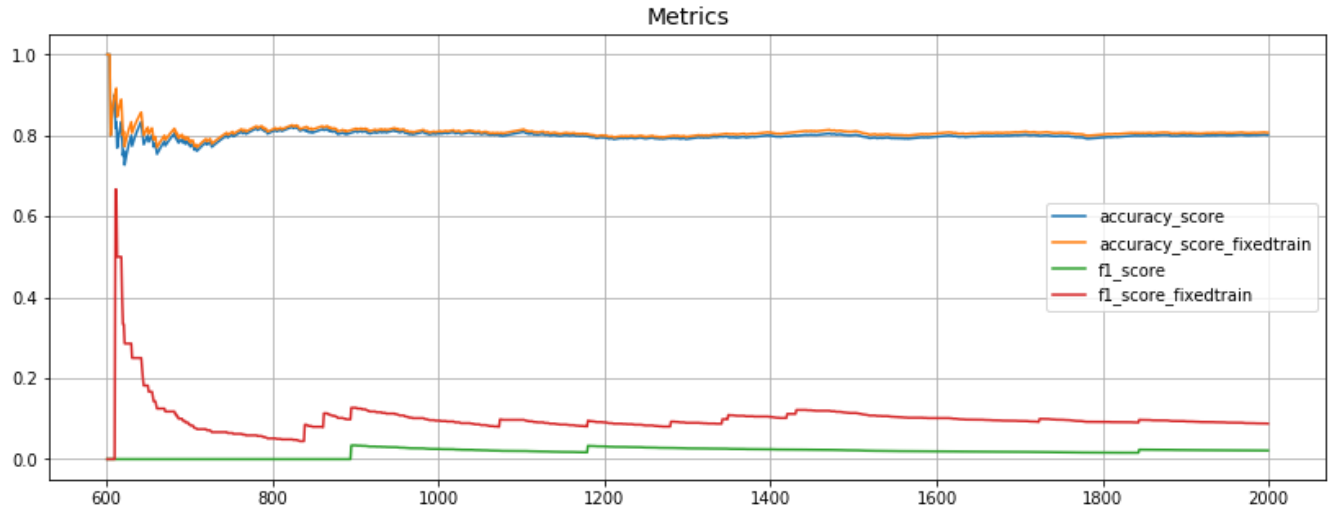
Improving time complexity

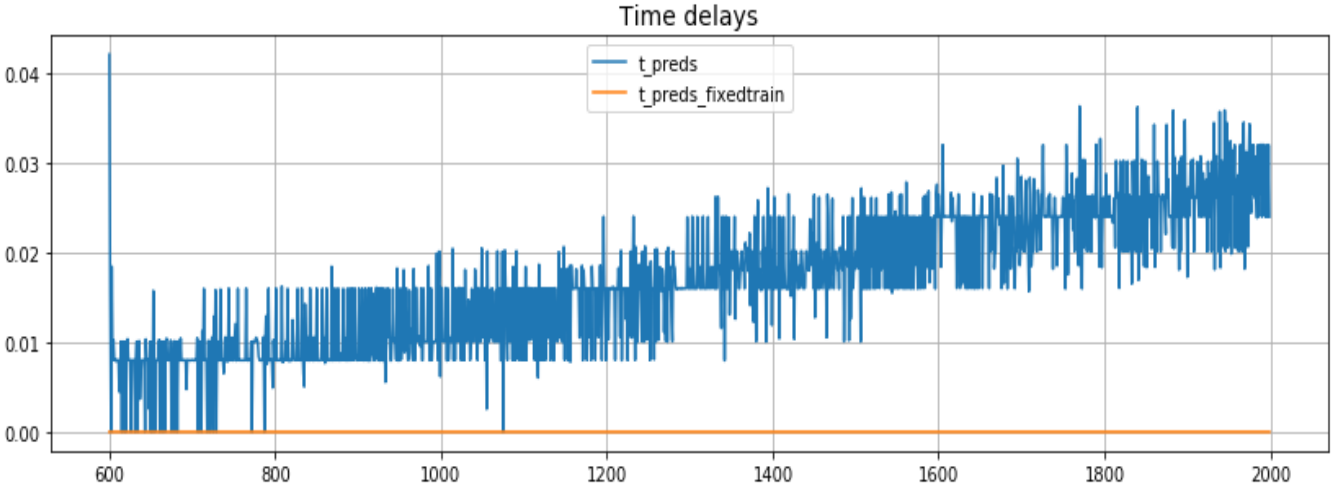
	precision	recall	f1-score	support
False	0.80	1.00	0.89	1119
True	1.00	0.01	0.02	281
accuracy			0.80	1400
macro avg	0.90	0.51	0.46	1400
weighted avg	0.84	0.80	0.72	1400

roc auc score: 0.5053

- Optimization:
  the search for counterexamples by intersection not with all the feaures, but only with those that were in intersection with positive examples.

- After the implementation of optimization:
  the execution time was reduced to 23 seconds.





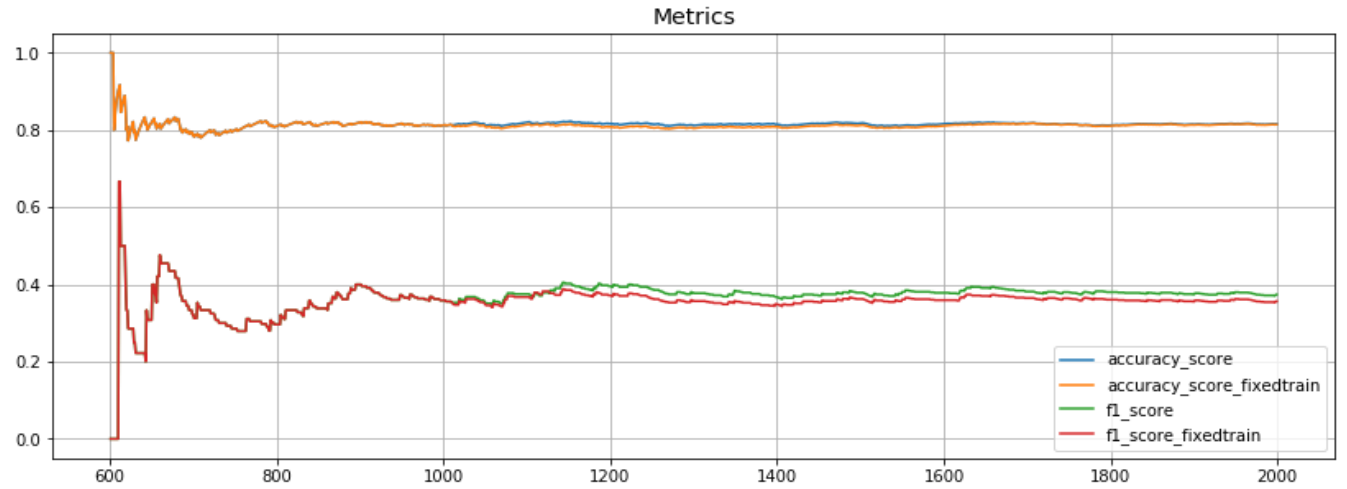
Modified algorithm

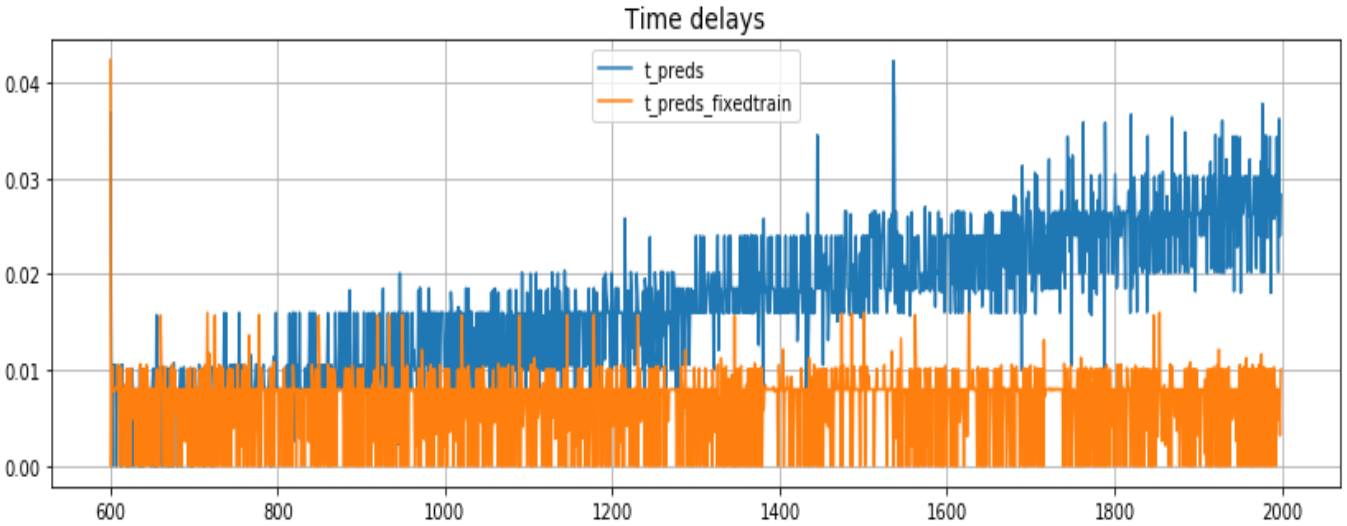
	precision	recall	f1-score	support
False	0.84	0.95	0.89	1119
True	0.59	0.27	0.37	281
accuracy			0.82	1400
macro avg	0.71	0.61	0.63	1400
weighted avg	0.79	0.82	0.79	1400

roc auc score: 0.8004

- Modification of the algorithm:
  to return as an answer not where there are more counterexamples, but the ratio of the normalized number of counterexamples.
  After such modification, the quality metrics have grown several times.

$$Ratio = \frac{counters_{neg}/N_{neg}}{counters_{pos}/N_{pos}}$$





Rule-based models

1. Catboost

	precision	recall	f1-score	support
0	0.86	0.96	0.91	1119
1	0.69	0.38	0.49	281
accuracy			0.84	1400
macro avg	0.77	0.67	0.70	1400
weighted avg	0.83	0.84	0.82	1400

roc_auc_score: 0.8252

2. Random Forest

	precision	recall	f1-score	support
0	0.85	0.97	0.91	1119
1	0.73	0.32	0.45	281
accuracy			0.84	1400
macro avg	0.79	0.65	0.68	1400
weighted avg	0.83	0.84	0.82	1400

roc_auc_score: 0.8241

- The quality of the two algorithms is comparable.
- The execution time of both algorithms is less than 1 second.
- ROC AUC is higher than the modified algorithm by 3%
- F1 score is higher than the modified algorithm by 27%

Thank you for your attention!