Customers churn prediction

Yevgen Shnurko
Big Data Lab 5
Vodafone

Customers churn

- Decreased revenue
- Increased costs
- Damaged reputation
- Loss of competitive advantage

Usually, it is more expensive to attract new customers than to retain existing ones.

Churn of 1000 customers:

1,440,000 ₴ per year

Decrease in profit *

300,000 ₴

Cost of acquiring the same number of new customers **

* Based on an average check of 120 ₴ per month

** Based on the cost of attracting a new subscriber 300 ₴

Project goal:

- Identifying customers who may leave the company..

For the sake of:

- Timely and effective retention measures. In particular, through special offers, personalized campaigns, service improvements, etc.

The 1000 subscribers that are pre-identified as churning is:

$$(1{,}440{,}000 \cdot K)$$ ₴ per year

By which we reduce the loss of income.

Where K is conversion, a measure of communication effectiveness.

$$0 \le K \le 1$$

Data:

- Main dataset: Monthly subscriber activity snapshot, 150,000 observations, 817 features.
- Additional dataset B_NUM: data on short numbers that interacted with subscribers, 671,248 observations, 8 features.
- Additional dataset DPI: subscriber mobile app traffic, 6,745,887 observations, 6 features.

Distribution of values of the target variable::

Training dataset:		Test dataset:	
0	140 414	0	140597
1	9 586	1	9403

All variables take exclusively numerical values.

201 variables have missing values

Uninformative variables:

- 11 features that do not take any value;
- 33 features that have only one unique value and no missing values;
- 2 features that have 2 and 6 non-empty observations, respectively, distributed across both classes;
- Subscriber ID.

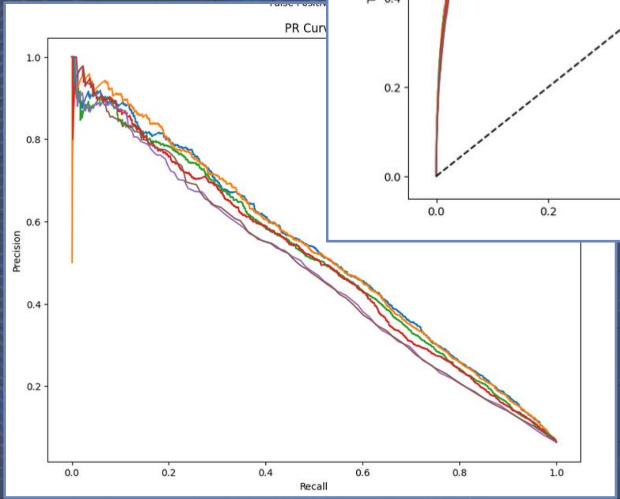
Models:

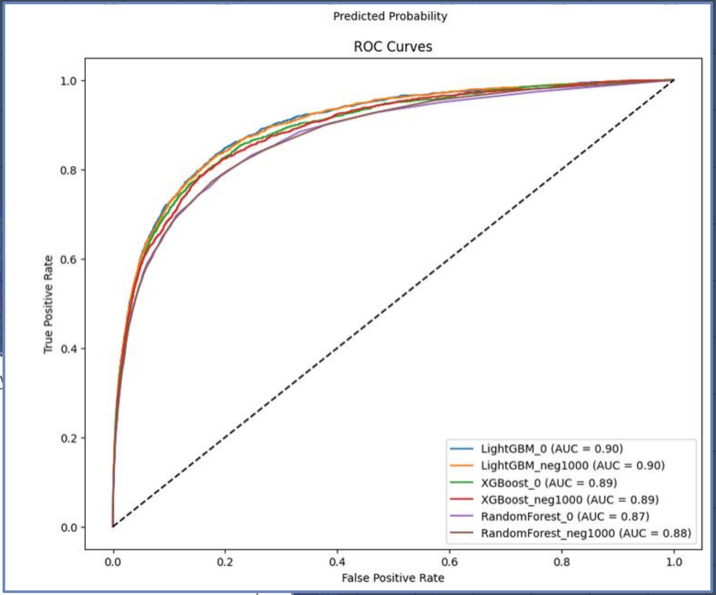
- LightGBM (Light Gradient Boosting Machine)
- XGBoost (Extreme Gradient Boosting)
- Random Forest

Training and validation were performed on a training dataset split 80 to 20.

Final testing of the model was performed on a test dataset.

Models:





AMD Ryzen 5 5500U with Radeon Graphics, 2.10 GHz

	LightGBM	XGBoost	Random Forest
Training Time, sec:	18.36	37.20	386.63
Prediction Time , sec :	3.80	2.49	25.96

Initial results:

- We leave the missing values as is;
- The model shows the best results on the top 25, 241 and 454 most important features.

Class balancing increases target group reach, but makes the issue of customers being misidentified as churn more pressing.

Additional datasets:

BNUM:

- 2301 additional features that characterize the subscriber's interaction with each of the short numbers.

DPI:

- 5 additional features that characterize the subscriber's interaction with applications in general;
- 2976 additional features that characterize the subscriber's interaction with each application, 744 features for each of the 4 metrics.

Best results:

- 25 best features of the main dataset and 744 features from the DPI dataset;
- The 40 most important ones were left, which included 25 features of the main dataset and 15 generated ones;
- The optimal model parameters were selected for the found combination of features.

```
params = {
    'boosting_type': 'gbdt',
    'num_leaves': 24,
    'max_depth': 6,
    'learning_rate': 0.010606
    'n_estimators': 985,
    'reg_alpha': 9.420645904
    'reg_lambda': 0.101262332
    'min_split_gain': 0.09655
    'subsample': 0.993145658
    'colsample_bytree': 0.766
    'objective': 'binary',
    'metric': 'auc',
    'is_unbalance': True
}
```

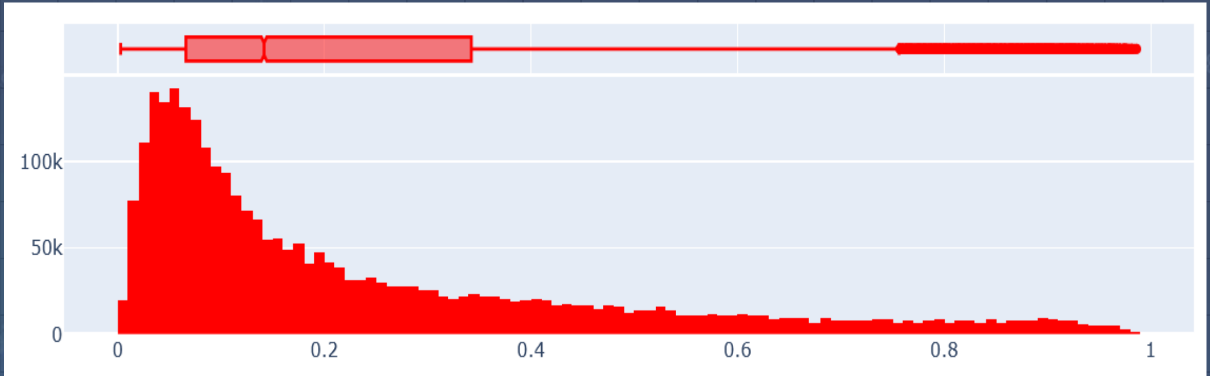
Model results:

TEST:

AUC: 0.8971

Recall: 0.7162

FP/TP Ratio: 1.9868



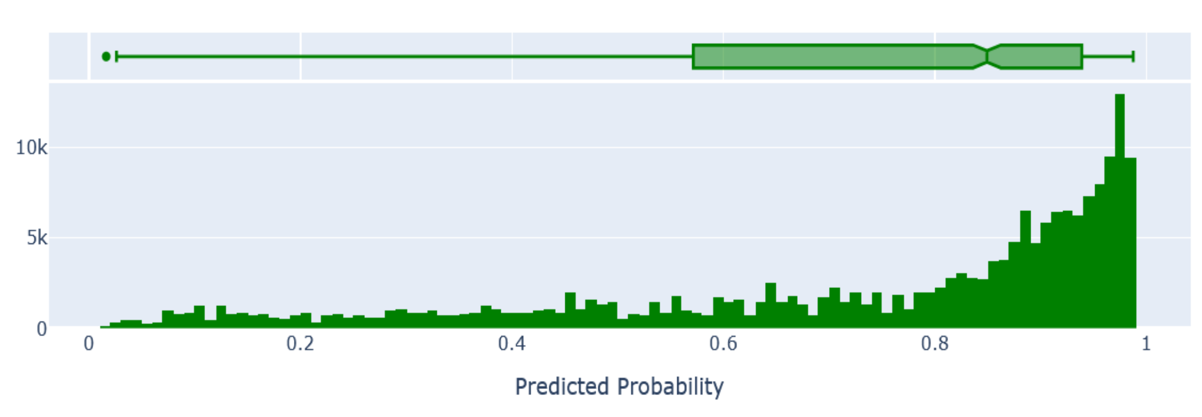
TRAIN:

AUC Train: 0.9302

AUC Validation: 0.9019

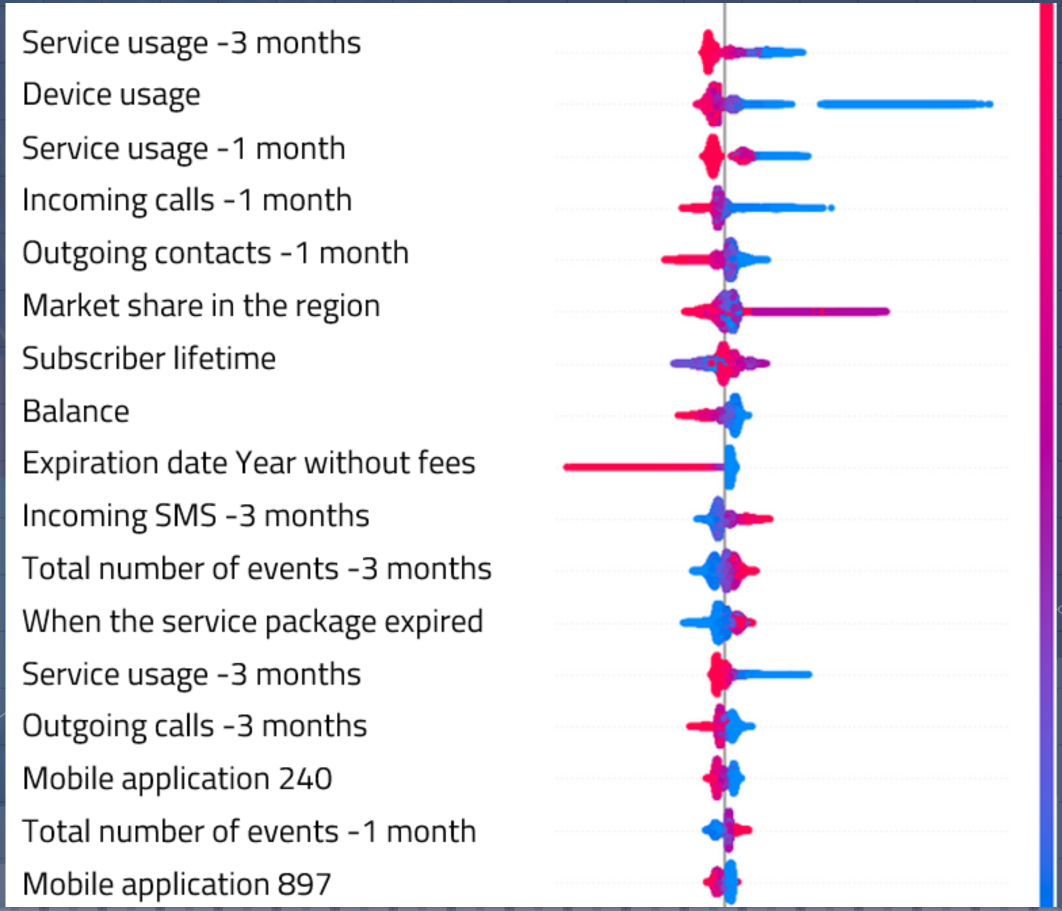
Recall: 0.7136

FP/TP Ratio: 1.9094

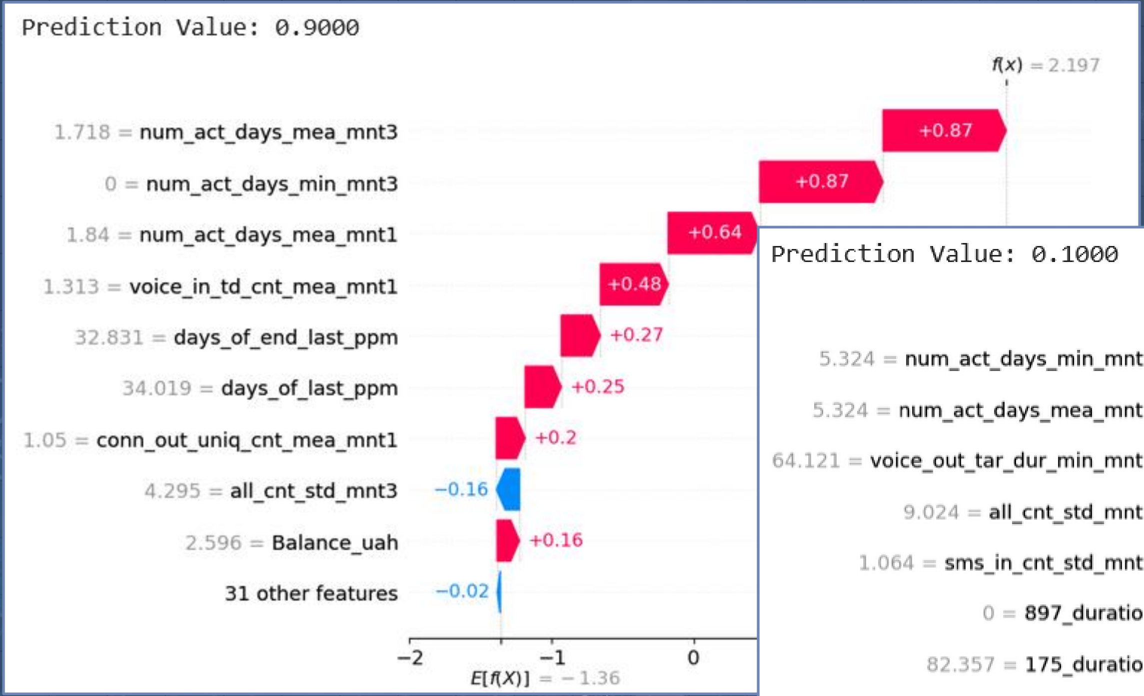


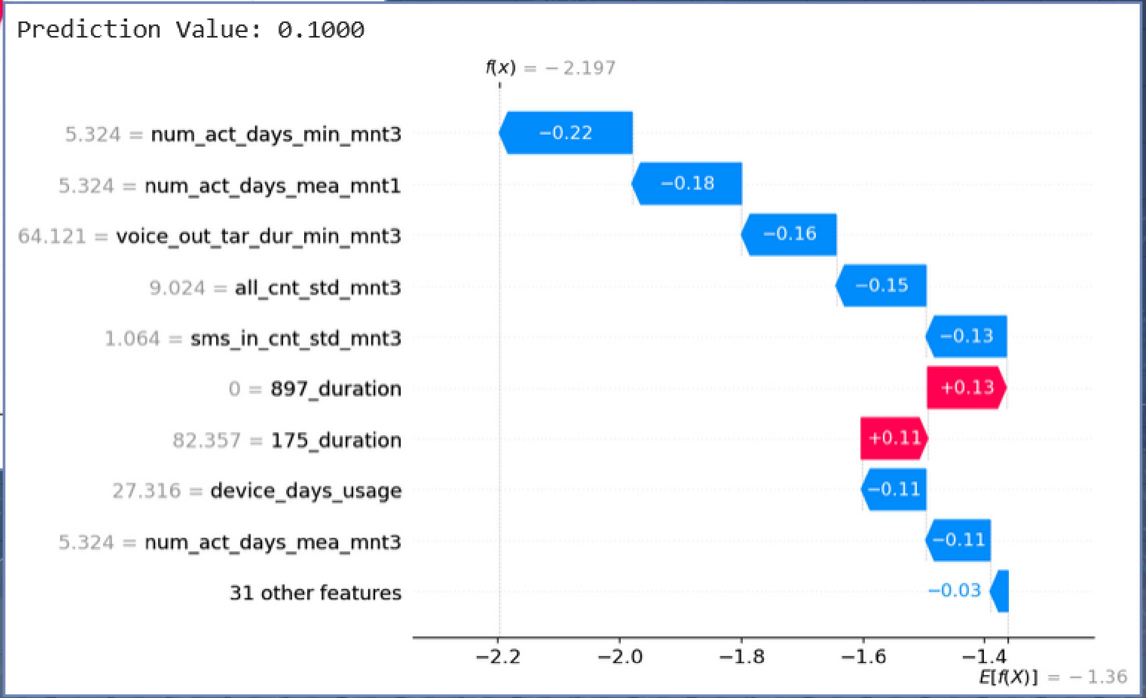
Interpretation of results

- Reducing service consumption
- Device usage
- Certain mobile applications

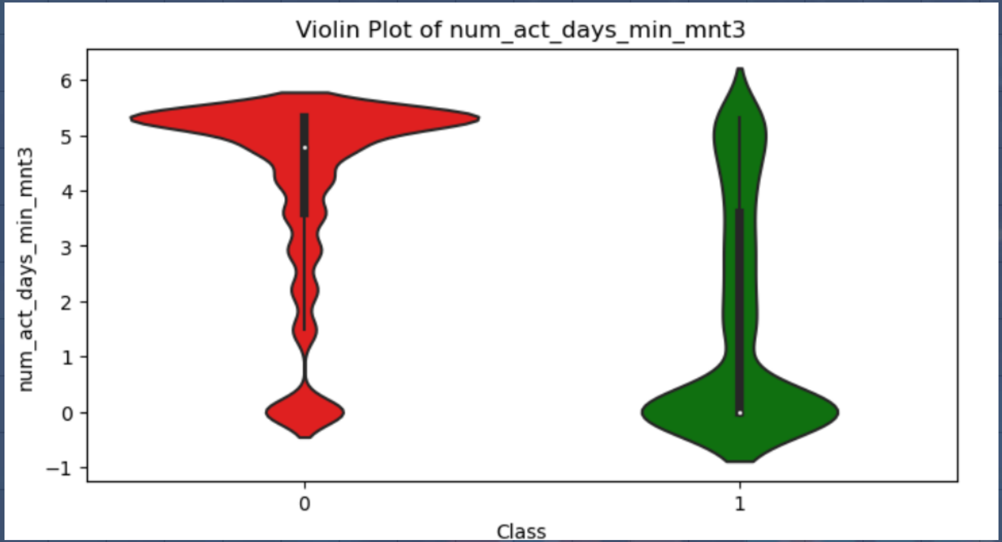


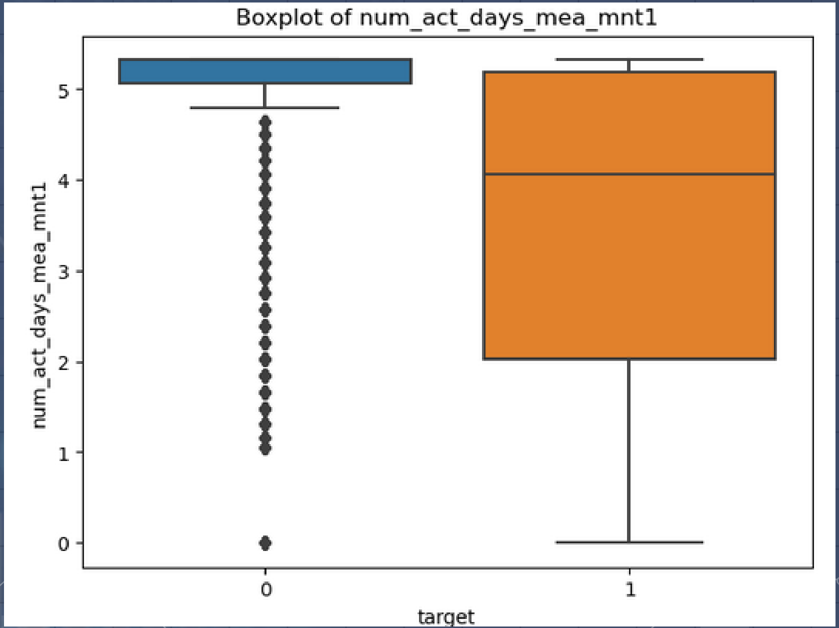
Interpretation of results





Reducing consumption





It is observed in both signs related to the period of -3 months and -1 month.

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

$$0,5 \cdot 120₴ \cdot 12 \text{ months} \cdot \text{Number of remaining}$$

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

$$0{,}5 \cdot 120₴ \cdot 12 \text{ months} \cdot \text{Correctly recognized} \cdot K_1 =$$

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

$$0{,}5 \cdot 120₴ \cdot 12 \text{ months} \cdot \text{Correctly recognized} \cdot K_1 =$$
$$= 0{,}5 \cdot 120₴ \cdot 12 \text{ months} \cdot \text{Misidentified} \cdot K_0$$

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

$$\cancel{0{,}5 \cdot 120\text{₴} \cdot 12\ \text{months} \cdot}\ \text{Correctly recognized} \cdot K_1 =$$
$$= \cancel{0{,}5 \cdot 120\text{₴} \cdot 12\ \text{months}} \cdot \text{Misidentified} \cdot K_0$$

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

For example, if we offer a 50% discount:

$$\frac{\texttt{Misidentified}}{\texttt{Correctly recognized}} = \frac{K_1}{K_0}$$

Recall vs False Positives:

How many falsely identified as at-risk can we afford?

Or in a more general case:

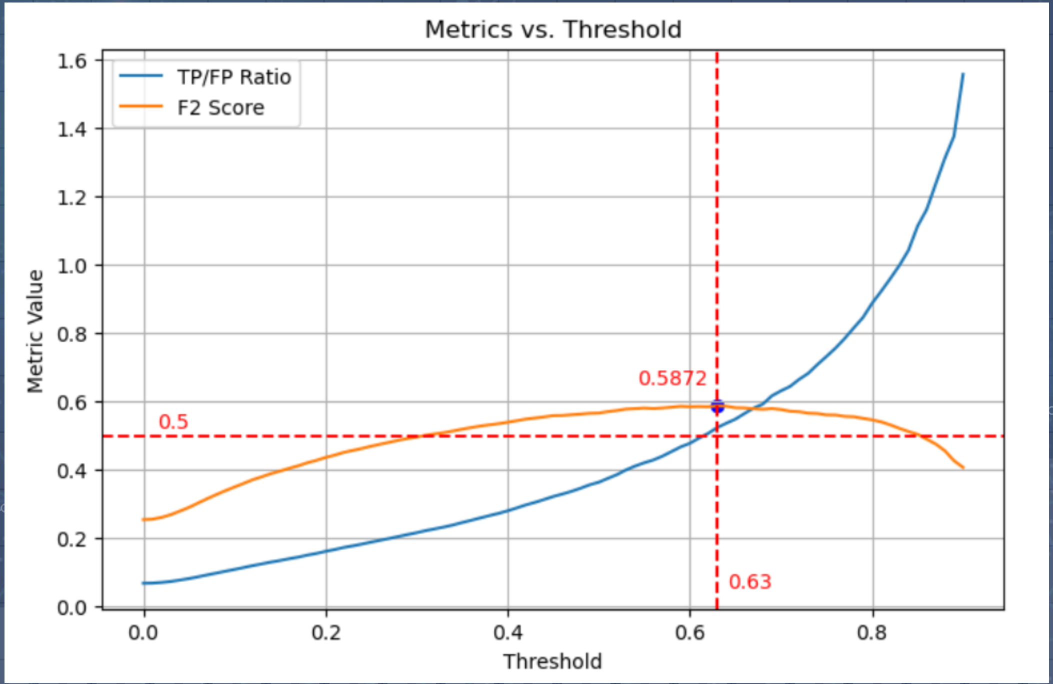
$$\frac{\text{Misidentified}}{\text{Correctly recognized}} = \frac{K_1}{K_0} \cdot \frac{\text{Discount, \%}}{(1 - \text{Discount, \%})} = R$$

The profit and cost ratio is added to the conversion ratio.

Recall vs False Positives:

The desired ratio is achieved by changing the threshold.

Maximum
$F_2 = 0,59$
Corresponding
$R = 2$ ($1/R = 0,5$)
at the threshold 0,63



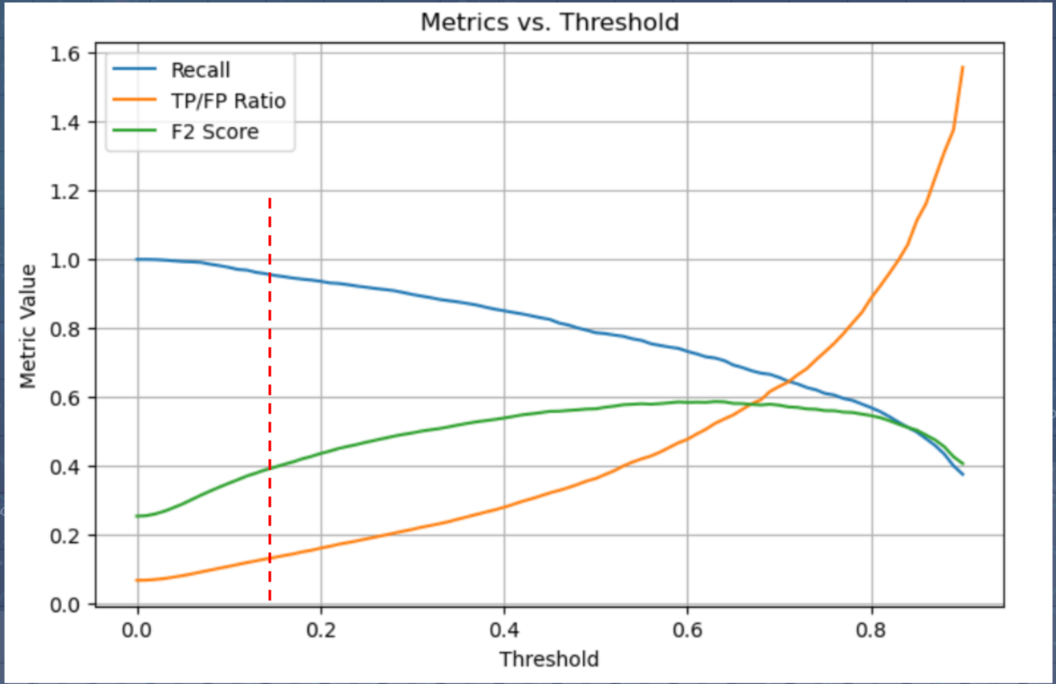
Recall vs False Positives:

With twice the conversion in the target group and a profit-to-cost ratio of 3 to 1:

$R = 6$

$F_2 = 0,4$

Recall = 0.93



Conversion calculation

Test sample from previous misidentified

May **June** **July**

K_0: conversion rate for falsely recognized

Current communication:

$R = FP / TP$

K: conversion for all recognized

K_1 - ?

Conversion calculation

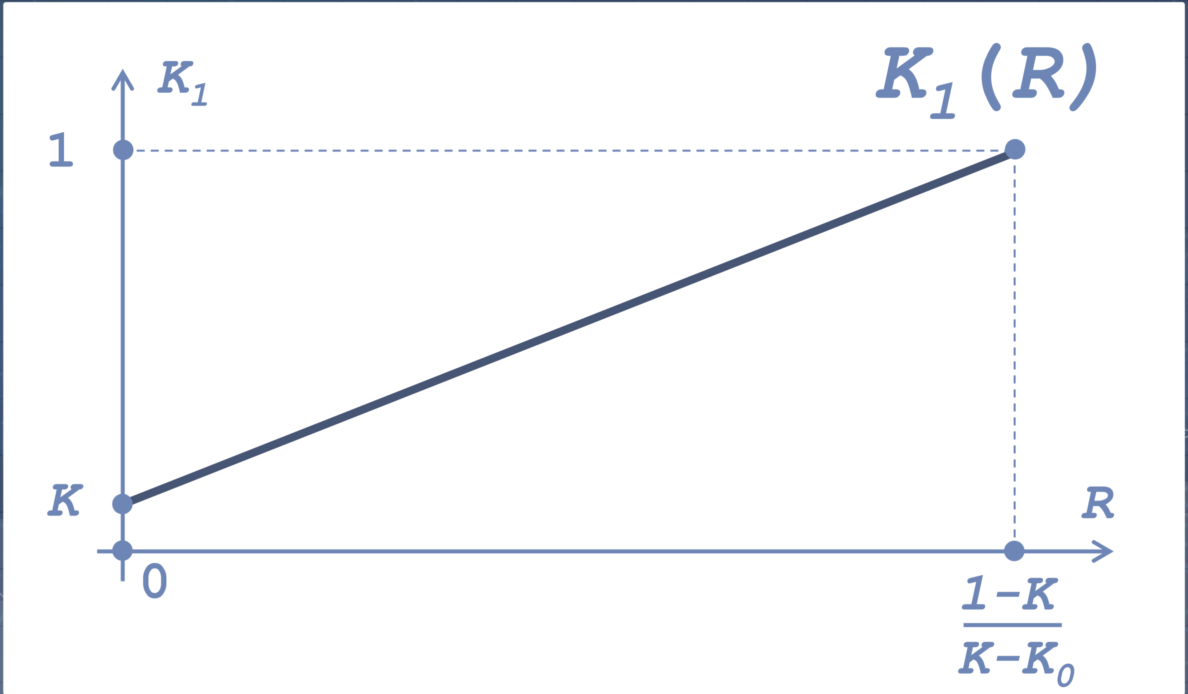
$$K_1 = R \cdot (K - K_0) + K$$

$K_1 > K$

$K > K_0$

$R = FP/TP$

$$K = \frac{FP \cdot K_0 + TP \cdot K_1}{FP + TP}$$

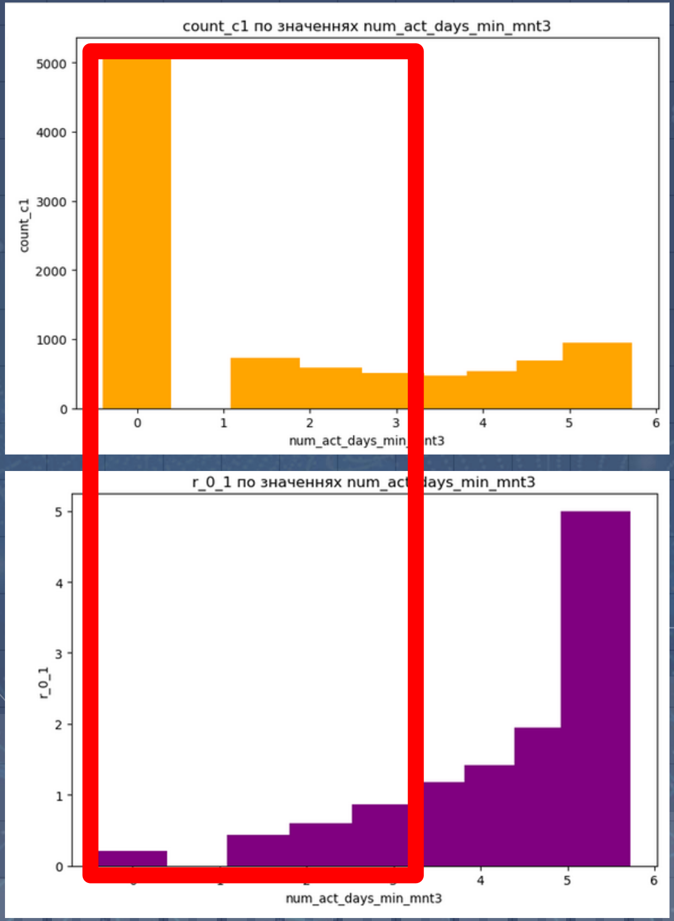


Data clustering

Number of days
of active use of services -3 months

Target group coverage: 72%
(for 4 values of the attribute out of 8)

AUC Test: 0.83
Recall Test: 0.6930
FP/TP Ratio Test: 0.84

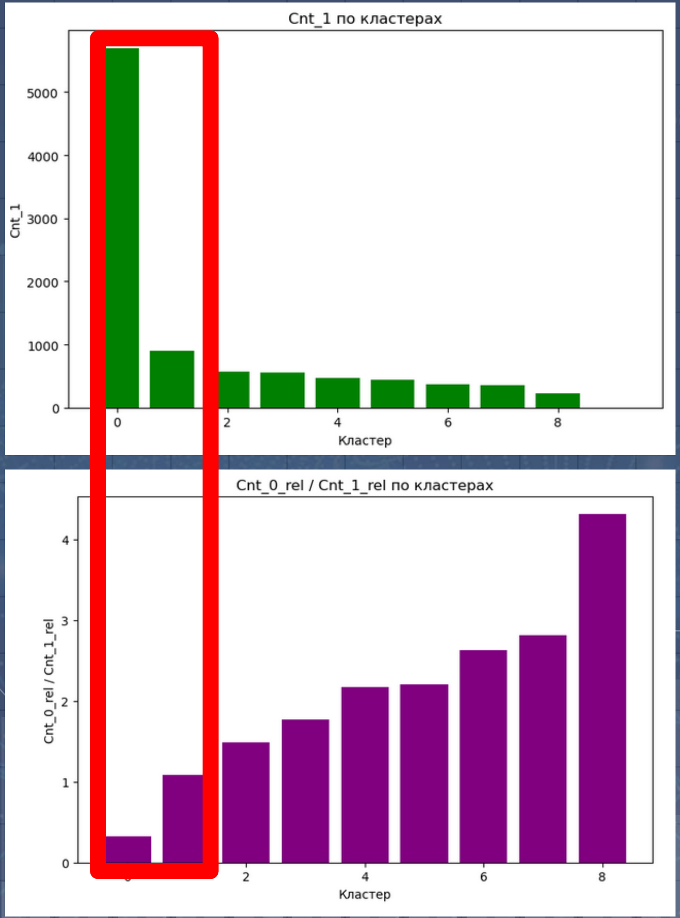


Data clustering

Number of unique outbound contacts –1 month

Target group coverage: 68%
(for 2 clusters out of 10)
AUC Test: 0.86
Recall Test: 0.72
FP/TP Ratio Test: 1.47

Target group coverage: 59%
(for 1 cluster out of 10)
AUC Test: 0.85
Recall Test: 0.70
FP/TP Ratio Test: 1.21



Data clustering

<u>Incoming SMS -3 months:</u> Coverage 51% (2 clusters out of 10)

AUC Test: 0.89 Recall Test: 0.76 FP/TP Ratio Test: 1.60

<u>Service Utilization (avg.):</u> Coverage 64% (2 clusters out of 10)

AUC Test: 0.83 Recall Test: 0.69 FP/TP Ratio Test: 1.07

<u>Number of events –3 months:</u> Coverage 51% (4 clusters out of 10)

AUC Test: 0.90 Recall Test: 0.76 FP/TP Ratio Test: 1.71

Conclusions

- Reducing service consumption requires immediate action;
- Communications should maximize conversion in the target group and minimize conversion for false positives;
- Use of mobile applications;
- None of the features related to technical problems entered the top 200 features in terms of importance;
- Competitor activity requires countermeasures;
- Importance of market share?

Thank you for your attention! Questions?








shnurko@gmail.com

+380 67 405 24 78

Images generated by the query "LightGBM predicts subscriber churn for Vodafone"