

The investigation of data temporal delays impact on classification performance for churn prediction in telecommunications

Andrej Bugajev, Rima Kriauzienė

Vilnius Gediminas Technical University

June 1

26th International Conference Mathematical Modelling and Analysis
MMA2023, May 30 — June 2, Jurmala, Latvia

The research practical importance

General importance

Currently in industry there is an emerging need for solution of problems which are considered out of scope of traditional mathematical modelling and currently is considered to be fallen into machine learning topic.

Customer behavior analysis

Customer activity often leaves data traces giving the opportunity for trials to reverse engineer reasoning behind the observed activity. Such type of analysis is among the ones which Machine learning techniques focused on.

Customer retention in telecommunications

According to other researchers, in the telecommunications industry acquiring a new subscriber costs 16 times more than retaining an existing one.

Churn definition in telecommunications

Often a churner in the mobile telecommunications is defined as a customer that stops doing revenue generating events during the next 90 days, while he was active during the observation period.

Some researchers use multiple definitions:

- full churner definition, the aforementioned definition based on 90 days absence,
 - partial churner definition with the shortened period of 30 days
-
- On the one hand, the shorter period lets to label more recent cases, thus it lets for a model to achieve a faster reaction to the changes in the behavioral patterns.
 - On the other hand, the quality of the possible prediction using shorter time interval might decrease due to possible labeling errors.

In previous research¹ we found out that with our data good compromise for Churn definition is 40 days of user absence.

¹Bugajev, A., Kriauzienė, R., Vasilecas, O., Chadyšas, V. (2022). The Impact of Churn Labelling Rules on Churn Prediction in Telecommunications. *Informatica*, 33(2), 247-277.

Customer churn prediction

An important role in success of customer retention plays the correct timing, there is a time window during which customer has more flexibility to change decisions. Thus, early churn prediction is the key factor in customer retention strategy.

Churn prediction problem

Churn prediction problem usually is formalized as classification problem. More specifically, for a chosen moment t using the last data before that moment the method must decide whether the customer will churn after moment t or not, i.e. whether he must be classified as churning.

$$\bar{Y}(X, H) = \left[\bar{y}_i(x_i) = \begin{cases} 1, & \text{if } \text{churner}(x_i, H) \\ 0, & \text{otherwise} \end{cases}, i \in [1, \dots, N] \right] \quad (1)$$

where $X = [x_1, x_2, \dots, x_N]$ is data, *churner* is some prediction method with parameters H . Then the problem can be formulated as optimisation problem

$$\min_H M(\bar{Y}(X, H), Y), \quad (2)$$

here M is the selected metric for prediction success, Y is the answer.

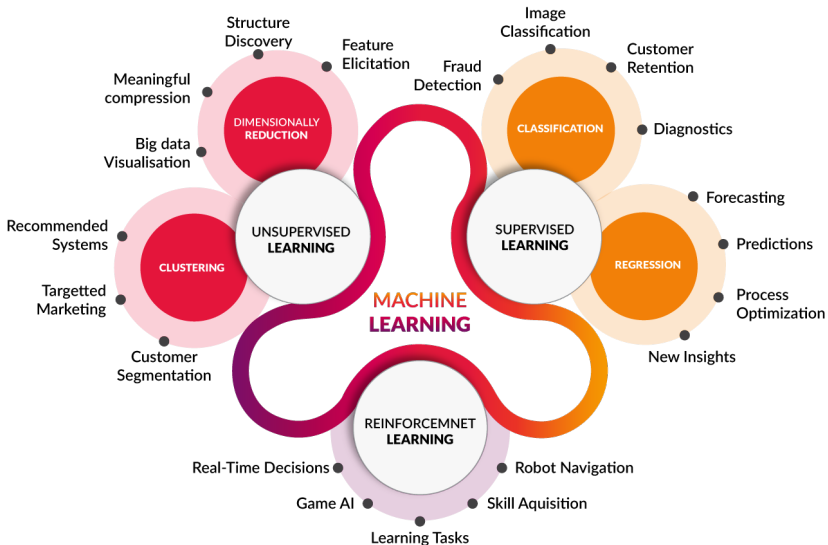


Figure: Machine learning overview. Source: <https://www.pngaaa.com/detail/3730478>

Table: Structure of the prepared dataset based on 90 days time interval

Attribute	Values or their range	Data type	Description
X1	0-26751	Numerical	The sum of minutes from all calls through whole period
X2	1-7032	Numerical	The amount of calls through whole period
X3	0-1475	Numerical	The sum of costs of customers payments through whole period
X4	0-255	Numerical	The amount of payments through whole period
X5	0-90	Numerical	The average of minutes from all calls during the day
X6	1-104	Numerical	Activity provided by company
X7	0-73	Numerical	Usefulness provided by company
X8	0-47	Numerical	Involvement provided by company
X9	0-266	Numerical	The maximum pause in days of customer activity
X10	0-275	Numerical	Duration of activities in days
X11, ..., X17	-	Numerical	The amounts of calls of different types (7 different)
X18, ..., X65	-	Numerical	RFM features
X66, ..., X425	-	Numerical	daily parameters for the last 90 days

Imbalance problem

Classification problem becomes imbalanced when there is an unequal distribution of classes in the dataset, this must be carefully addressed. The solution might converge to unexpected result, the metrics might become uninformative.

The Churn rate in our dataset is 20.21%, which means, the imbalance in the considered data is not very high, this will be taken into account later.

Solution:

- use balancing techniques for the training part of the dataset, such as undersampling and oversampling – we will use random oversampling;
- use appropriate metrics constructed for imbalanced data.

Next, we will discuss different metrics in the context of the churn problem.

Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

The standard notations of TP, FP, FN, TN will be used, which are defined by confusion matrix.

Accuracy is the most common metric to measure the performance of classification method. What it lacks is the sensitivity to data imbalance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall

Recall or true positive rate (TPR) is calculated as $TPR = TP / (TP + FN)$. In the context of churn prediction this is an important metric, i.e. we want to be sure that if a customer is going to leave then a retention technique should be applied, even if it will lead to a side effect of applying the retention to some of non-churners falsely classified as churners.

Precision

Precision or positive predictive value (PPV) is calculated as $PPV = TP / (TP + FP)$, it shows how well the positive result is determined. The indirect usage of it might be useful, for example as a criteria to limit the recall.

F1 score

The F1 score is a commonly used performance metric for imbalanced data cases. F1 score combines precision and recall into a single value. F1 score is the harmonic mean of precision and recall

$$F_1 = 2 \cdot PPV \cdot TPR / (PPV + TPR).$$

Here we will discuss time windows for data collection, according to single month based churn definition.

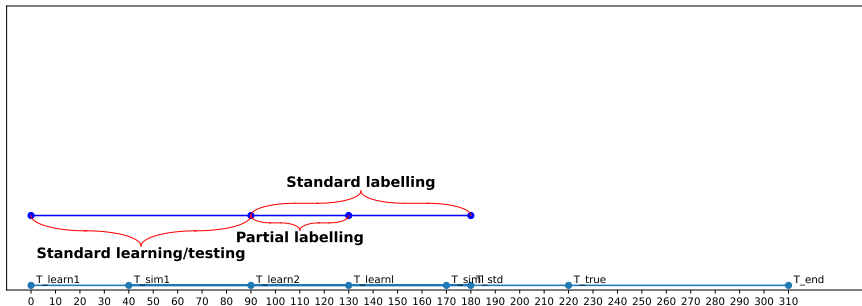


Figure: Time windows for constructing the datasets and setup the experiment.

The main research question – how do time delays affect the performance of prediction models?

Here we will discuss time windows for data collection, according to single month based churn definition.

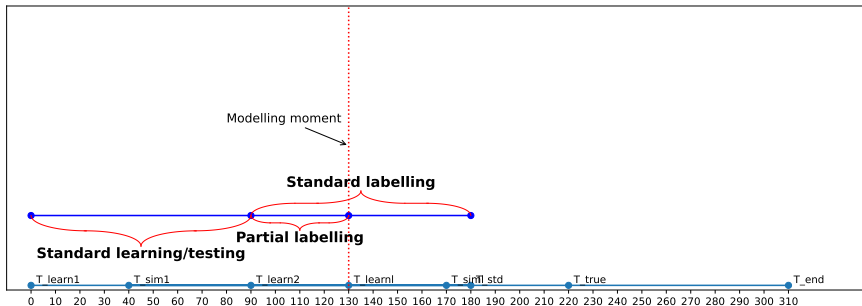


Figure: Time windows for constructing the datasets and setup the experiment.

The main research question – how do time delays affect the performance of prediction models?

Here we will discuss time windows for data collection, according to single month based churn definition.

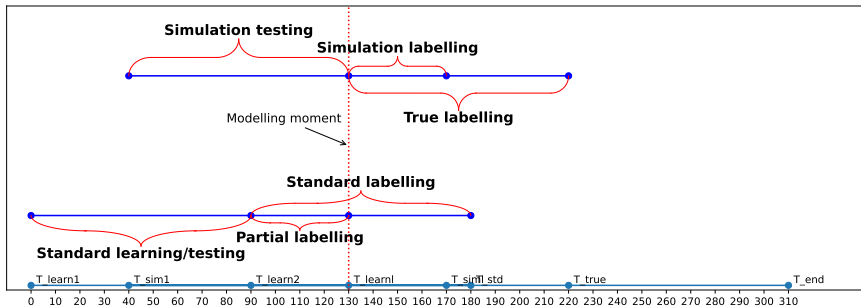


Figure: Time windows for constructing the datasets and setup the experiment.

The main research question – how do time delays affect the performance of prediction models?

Here we will discuss time windows for data collection, according to single month based churn definition.

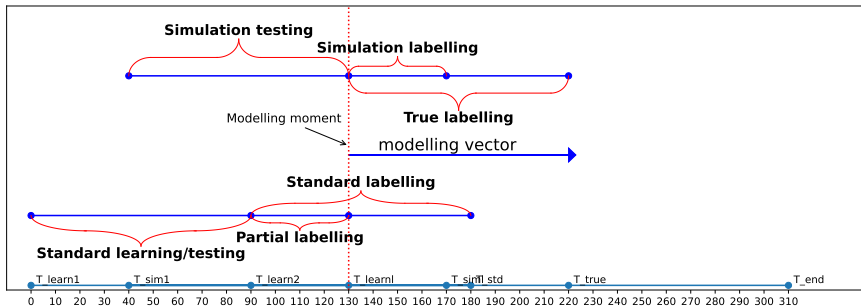


Figure: Time windows for constructing the datasets and setup the experiment.

The main research question – how do time delays affect the performance of prediction models?

The steps to perform the churn prediction are:

- 1 RFM and other features extraction, data labelling
- 2 fitting the unsupervised methods pipeline: feature vectors' normalisation by standard scaler (division by standard deviation), application of PCA (Principal component analysis),
- 3 the construction of classification method with the selected parameters,
- 4 fitting the model to the obtained data, model evaluation is applied to the separated validation part, and the training procedure is applied to the enriched training part, obtained using random oversampling technique.
- 5 the data for prediction is transformed using the same unsupervised methods pipeline, classification model is applied to the resulted data.

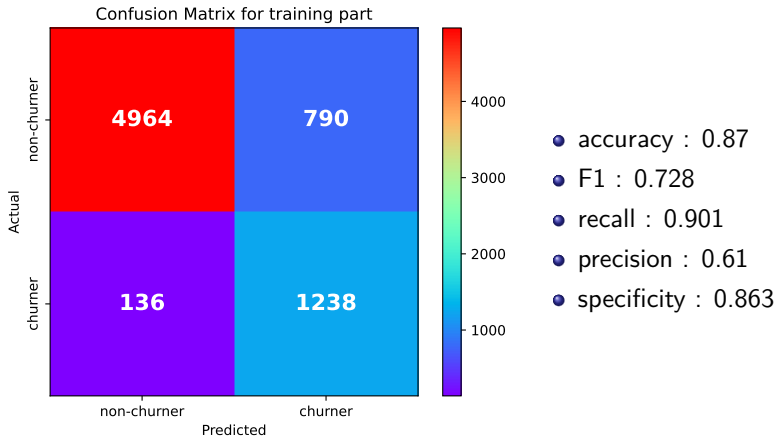
Hyperparameters ranges of classification methods

Two methods were analysed:

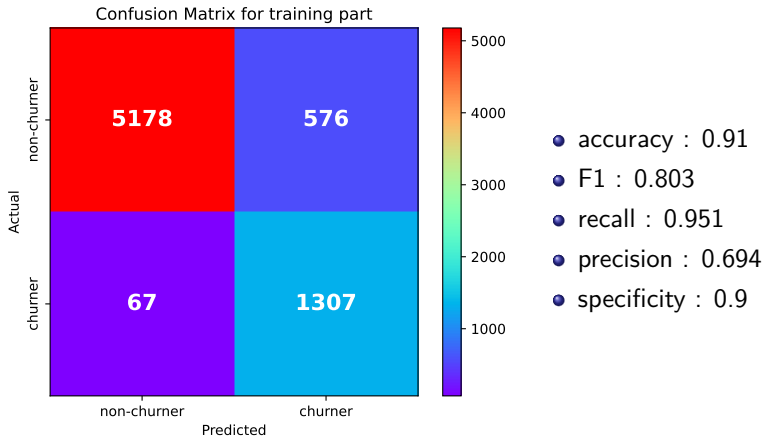
- 1 Gradient Boosting Classifier (CBC) from sklearn library,
- 2 CatBoost classifier – open source gradient boosting enhanced method actively developed by Yandex.

Method	Hyperparameters	Ranges
GBC	<i>min_samples_leaf</i> :	[3, 5, 7],
	<i>n_estimators</i> :	[256, 512],
	<i>max_depth</i> :	[2, 3, 5, 7],
	<i>n_iter_no_change</i> :	[50],
	<i>tol</i> :	[0.0001];
CBC	<i>min_child_samples</i> :	[3, 5, 7],
	<i>n_estimators</i> :	[256, 512],
	<i>max_depth</i> :	[2, 3, 5, 7],
	<i>early_stopping_rounds</i> :	[50],

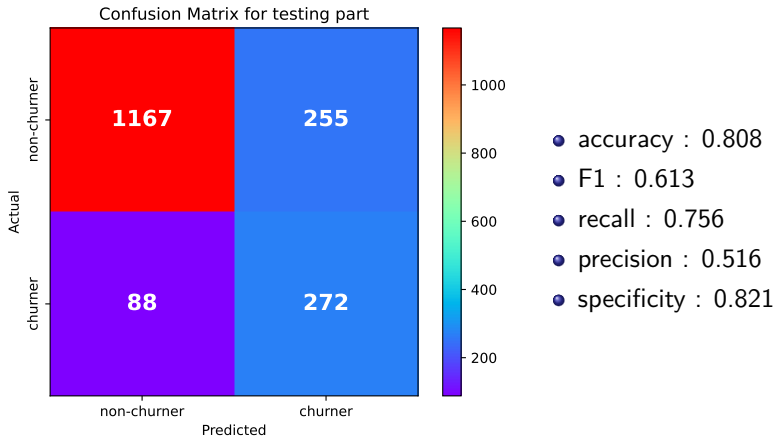
Training (GBC)



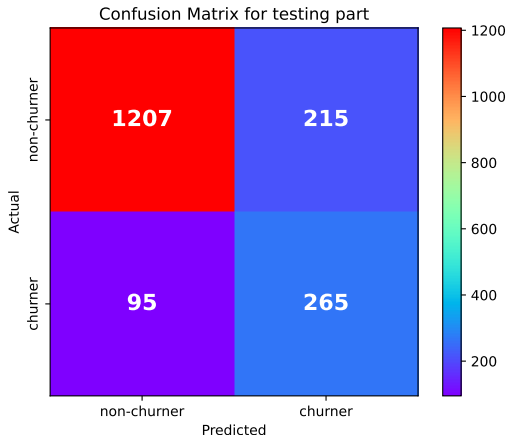
Training (CBC)



Testing (GBC)

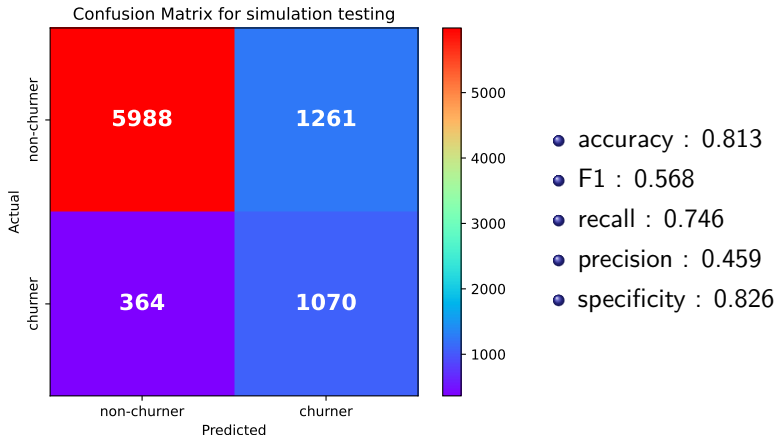


Testing (CBC)

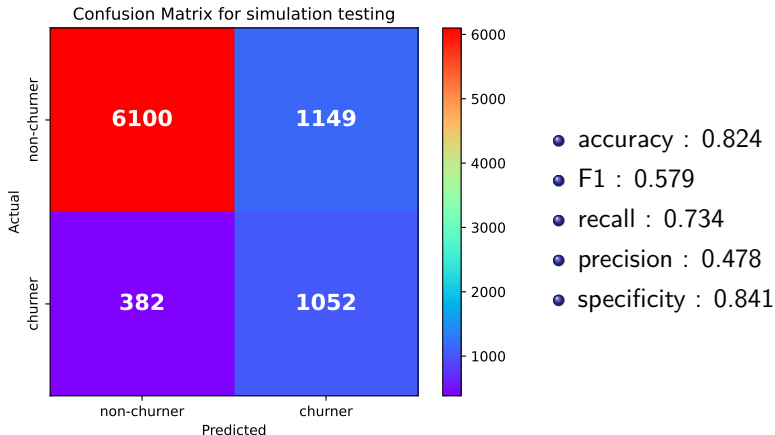


- accuracy : 0.826
- F1 : 0.631
- recall : 0.736
- precision : 0.552
- specificity : 0.849

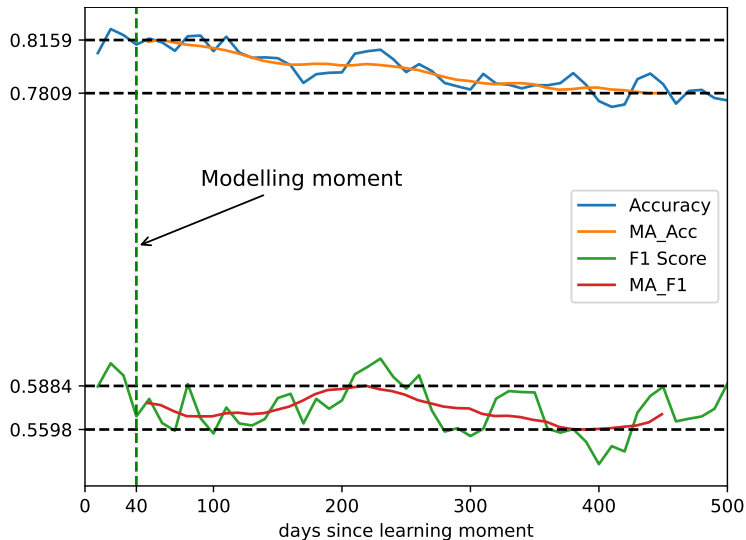
Simulation testing (GBC)



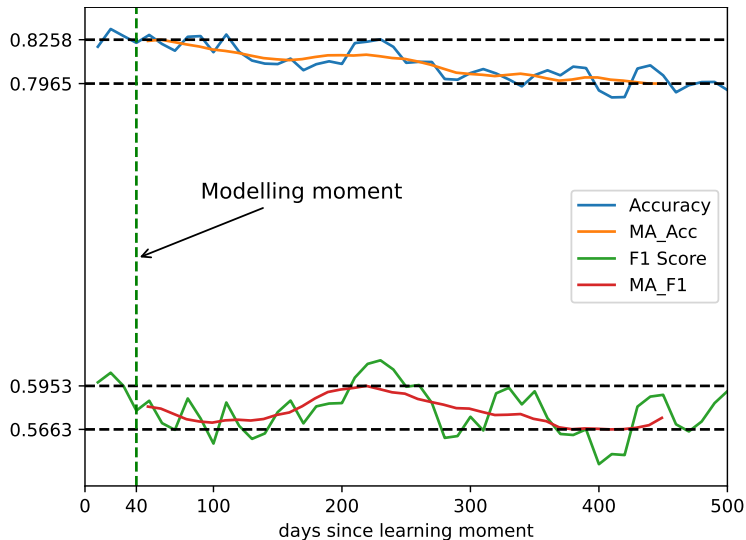
Simulation testing (CBC)



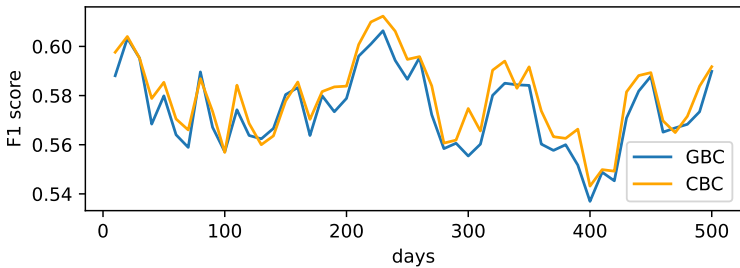
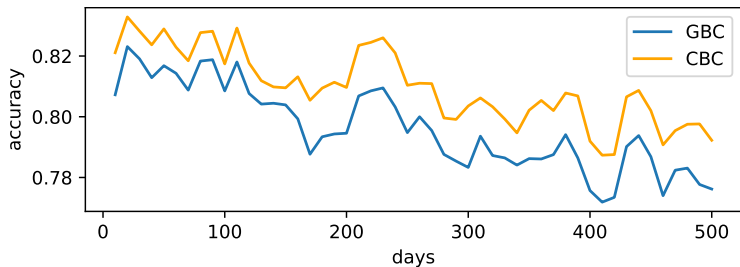
The results with delays using GBC



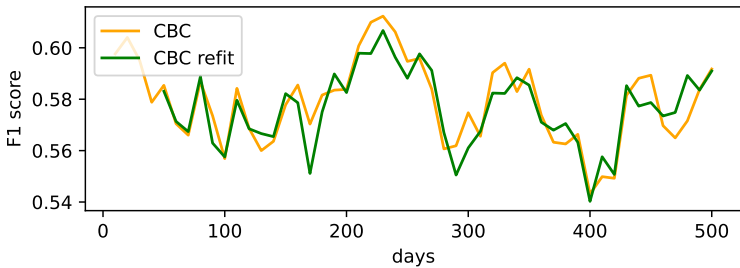
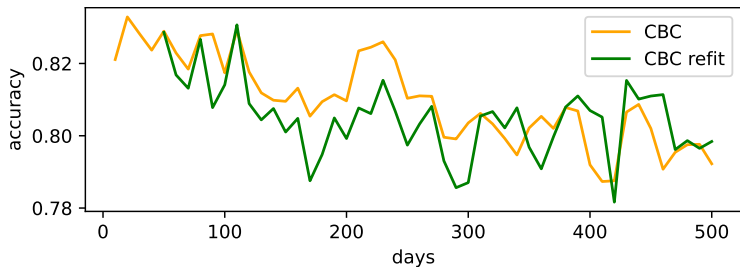
The results with delays using CBC



GBC vs CBC




GBC refit



- 1 CBC is a better alternative to GBC.
- 2 The simulation of model application shows that the mode is adapted to the learning moment, resulting in Accuracy drop from 0.826 to 0.824 and F1 score drop from 0.631 to 0.579. from
- 3 However, the model is very robust to the possible long term behavioural changes in time – after 500 days we were unable to identify any changes in the model's performance.
- 4 To improve the performance we recommend to include into training and/or validation data some of data which was involved in simulation testing.

Quality

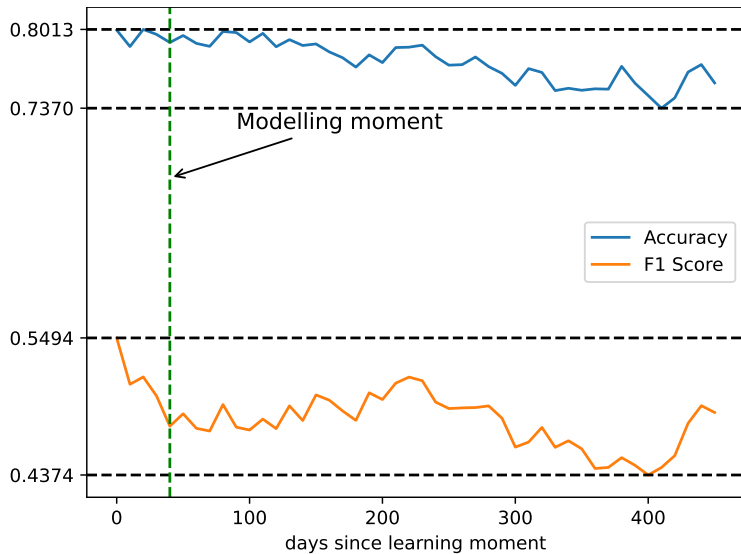
	CatBoost		LightGBM		XGBoost	
	Tuned	Default	Tuned	Default	Tuned	Default
 KDD churn	0.23129	0.23193 +0.28%	0.23205 +0.33%	0.23565 +1.89%	0.23312 +0.80%	0.23369 +1.04%

Learning speed

	CatBoost	LightGBM	XGBoost
CPU (Xeon E5-2660v4)	527 sec	1146 sec	4339 sec
GTX 1080Ti (11GB)	18 sec	110 sec	890 sec

Dataset Epsilon (400K samples, 2000 features). Parameters: 128 bins, 64 leafs, 400 iterations.

The results with true churn labels



The findings in this study raise other questions that might be considered as research gaps:

- Do companies actually need a binary classification of churners, if the result is sensitive to the assumptions that look natural? Some sort of alternative classification generalization could be considered. Especially it can be true since these days changing operators is easy, also new eSIM technology possibilities appeared, the loyalty to some services of companies might be much more fuzzy than it was a couple of decades ago.

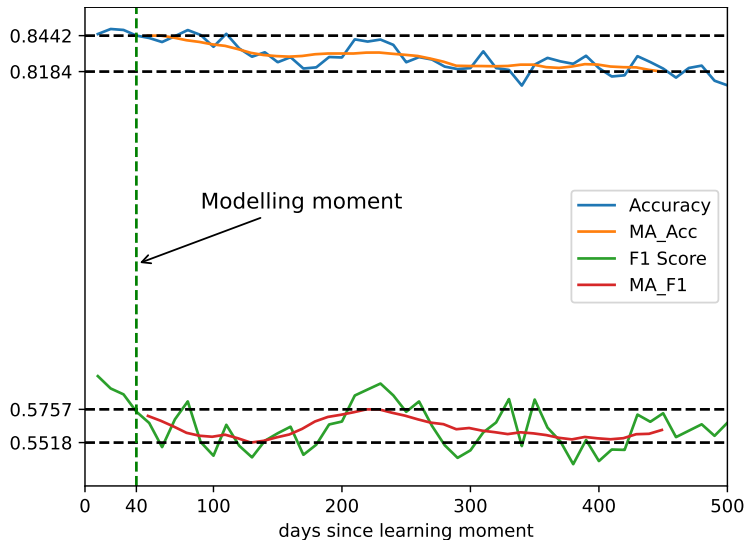
Specificity

Specificity or true negative rate (TNR) is calculated as $TNR = TN / (TN + FP)$. Must be paid attention, if we want to limit the amount of retention applied to non-churners, this characteristic measures that indirectly, since you can derive false positive rate from it $FPR = 1 - TNR$.

F1 score

The F1 score is a commonly used performance metric for imbalanced data cases. F1 score combines precision and recall into a single value. It provides a balanced measure of the model's accuracy by considering both the false positives and false negatives. Moreover, it is paid attention to the fact that we are talking about rates – the harmonic mean is used instead of arithmetic average. Thus, F1 score is the harmonic mean of precision and recall $F_1 = 2 \cdot PPV \cdot TPR / (PPV + TPR)$.

The results with different model application delays



Recency, Frequency and Monetary (RFM) features are well-known for their suitability for churn modelling. These features are the values obtained by aggregation of the considered parameter in the selected time interval in three different ways.

RFM features are derived from this daily data:

- The sum of minutes from all calls during the day
- The amount of calls during the day
- The sum of costs of customer payments during the day
- The amount of payments during the day

Let us assume that we have daily data, i.e. the sums of some parameter for different days, we denote t_1, t_2 to be integers representing the first and last day of investigated interval then R, F, M features in interval $t = [t_1, t_2]$ are obtained using these formulas:

$$R(t_1, t_2) = \begin{cases} \sum_{ts+1}^{t_2} 1, & \text{if } ts < t_2 \\ 0, & \text{otherwise.} \end{cases}, \quad (3)$$

$$ts = \max\{j : f_j > 0, f_j \in \{f_{t_1}, f_{t_1+1}, \dots, f_{t_2}\}\}$$

i.e. R (Recency) is equal to the amount of the last days with non-zero values of selected parameter during the interval $[t_1, t_2]$.

$$F(t_1, t_2) = \sum_{t=t_1}^{t_2} \begin{cases} 1, & \text{if } f_t > 0, \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

i.e. F (Frequency) is the amount of days with non-zero values of selected parameter.

$$M(t_1, t_2) = \sum_{t=t_1}^{t_2} f_t, \quad (5)$$

thus, M (monetary) is a basic sum, in other words it transforms daily data to some data aggregated in the same way, but for the bigger time interval.

The results with true churn labels

