Выбираем
CatBoost или XGBoost

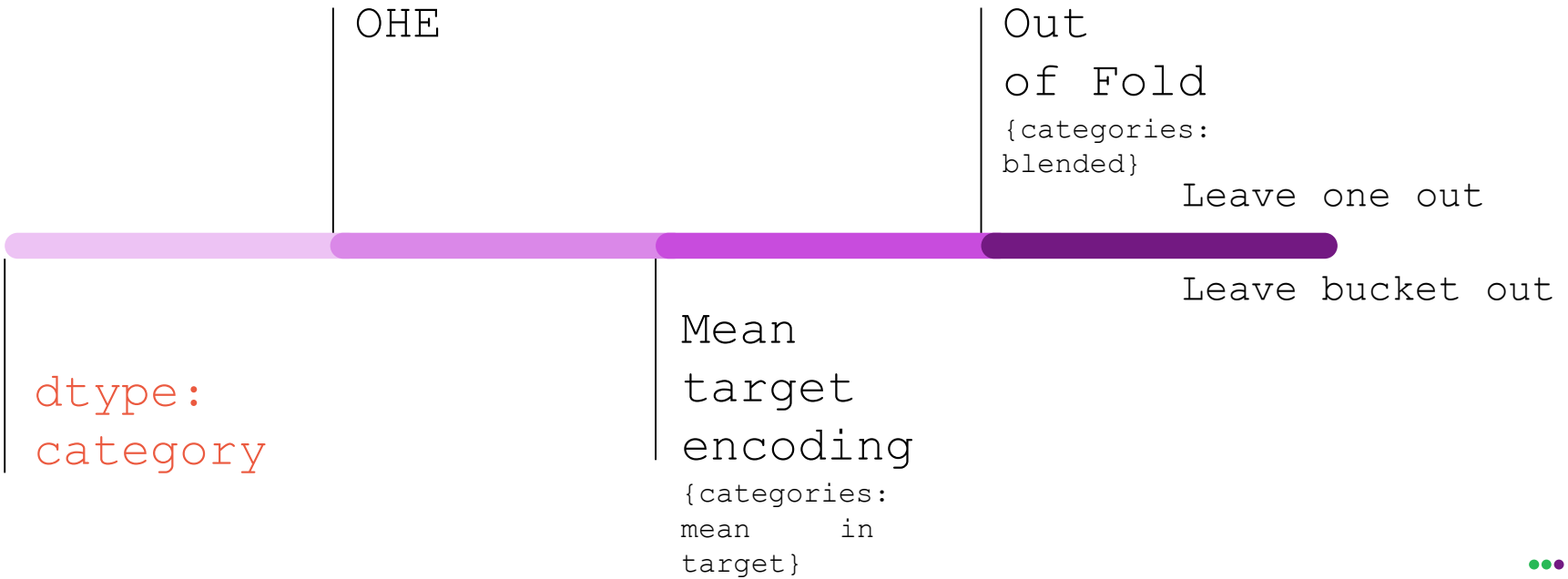
Селезнев Артем, 27 Апреля 2019

МегаФон

CatBoost?

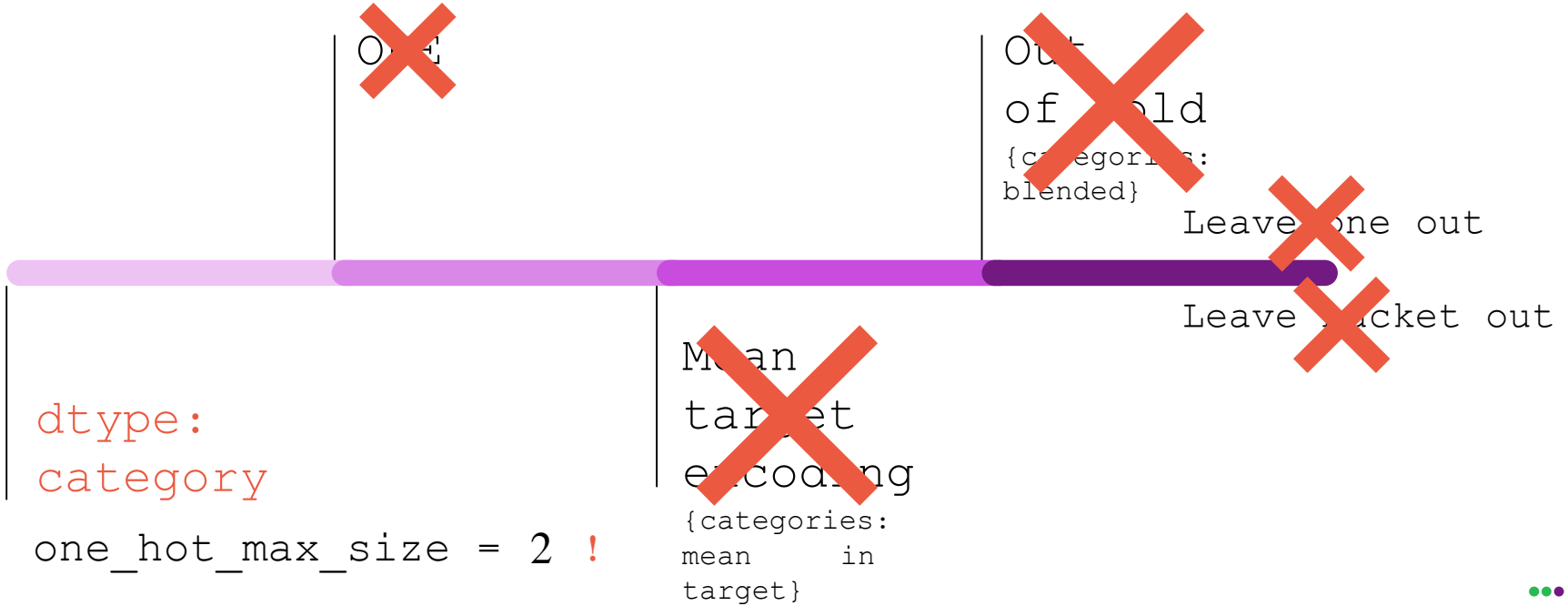
CatBoost?

1. “Cat” isn’t a pussy!



CatBoost?

1. “Cat” `isn’t` a pussy!



CatBoost?

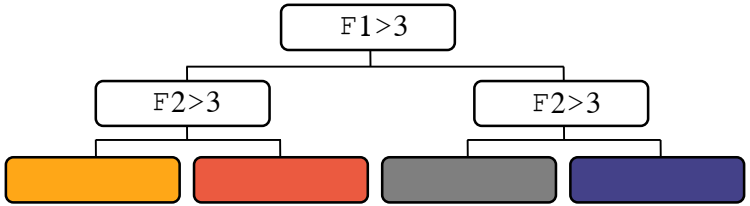
2. Oblivious Trees inside

CatBoost?

2. Oblivious Trees inside

Max_depth: 1 – 16

Особенность!
Best: 6, 10



CatBoost?

3. Бинаризация данных

XGBoost

1. Uniform

CatBoost

1. Uniform — CPU
2. Median
3. Quantile
4. MaxSumLog — Prod Y
5. GreadyLogSum

CatBoost?

4. Меньше параметров для подбора (влияющих на результат)
    1. learning_rate ⬇
    2. n_estimators ⬆
    3. overfitting detection settings + eval
    4. L2_leaf_reg
    5. bagging_temperature
    6. random_strength

CatBoost?

4. Меньше параметров для подбора (влияющих на результат)
   1. learning_rate ⬇
   2. n_estimators ⬆
   3. overfitting detection settings + eval
   4. L2_leaf_reg
   5. bagging_temperature
   6. random_strength

colsample_bylevel

CatBoost?

еще немного ...

5. Обучение из BaseLine (auto)
6. Eval_metrics (overfitting detection)
7. Staged_predict (поэтапное обучение)
   (! управляемое)
8. SnapShots

CatBoost пока не может

01

XGBoost spark

02

Иногда плох без GPU

03

Все ещё не быстрый

04

Надо явно указывать категориальные данные (cat_features)

TASK

DataSet (churn)

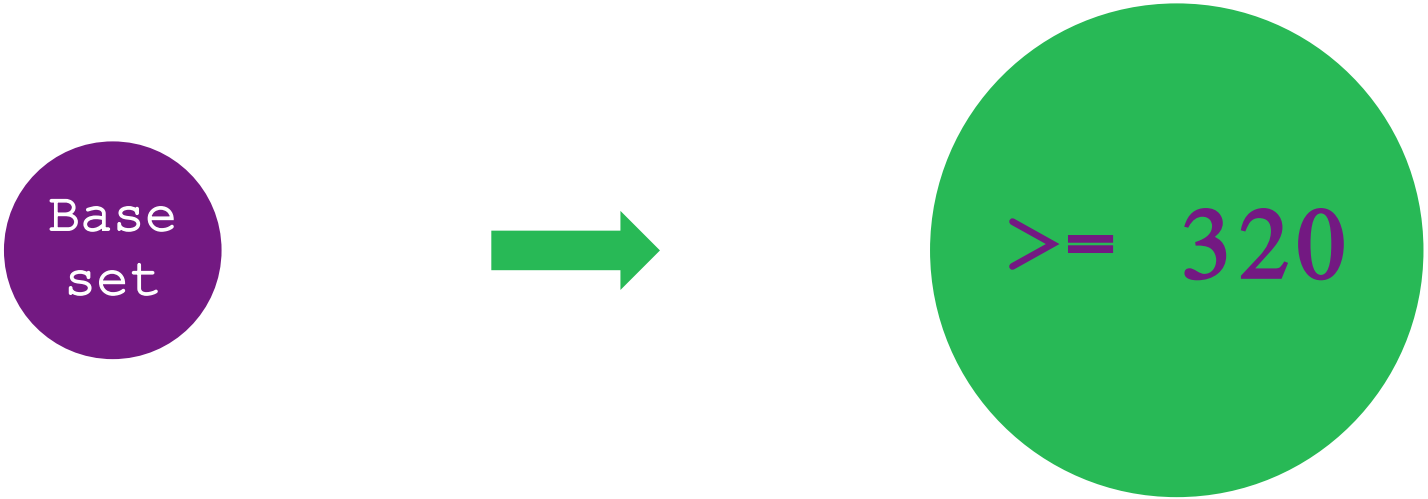
1. General_final – информация об абоненте (date_key, test > 13.05)
2. Traf_final – агрегированная информация о трафике
3. Rech_final – агрегированная информация о пополнениях
4. Balance_final – информация о балансе на конец дня

>= 320 шт

TARGET (churn + N months hold)

Преобразование

1. Изменение параметров значения:
   по отрезкам, накопит. итог, за период



А можно что-то изменить?

CatBoost VS XGBoost

А результат?

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

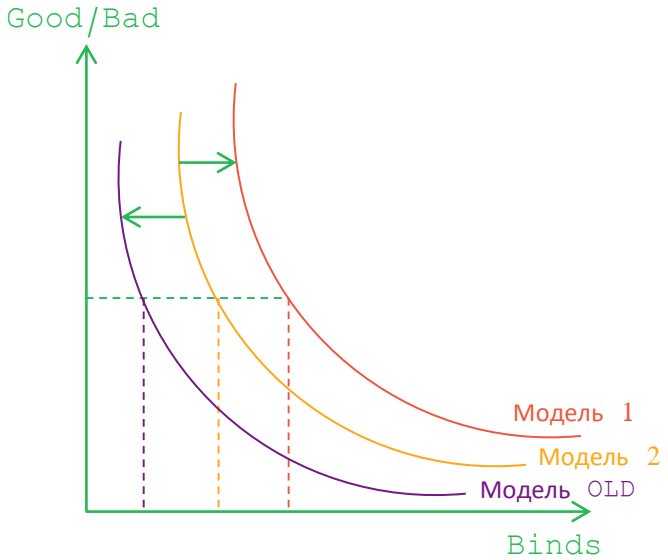
   1. Построение таблицы по каждой модели

bind_type	model 1	...	model №
Bind	Bad % / Good %	:-:	Bad % / Good %

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

   1. Построение таблицы по каждой модели
   2. График



Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

   1. Построение таблицы по каждой модели
   2. График
   3. Дивергенция

$$DIV = \frac{(\overline{good} - \overline{bad})^2}{[0.5 * (\sigma^2_{good} + \sigma^2_{bad}))]}$$

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

   1. Построение таблицы по каждой модели
   2. График
   3. Дивергенция
   4. Таблица выигрышей

	score group	количество в группе	факт.good	факт.bad	model good	model bad
Model Old						
Model 1						
.....						

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

   1. Построение таблицы по каждой модели
   2. График
   3. Дивергенция
   4. Таблица выигрышей
   5. Стабильность

	bind_type	actual %	expected %	actual - expected	actual / expected	ln(actual / expected)	index
Model Old							
Model 1							
.....							

index расчитывается по формуле:

$$index = \frac{\sum(actual\% - expected\%))}{\ln(actual\%/expected\%)}$$

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
   github.com/NameArtem/papers/blob/master/ML_model_comparison.md

   1. Построение таблицы по каждой модели
   2. График
   3. Дивергенция
   4. Таблица выигрышей
   5. Стабильность
   6. Групповой отчет

score	expected good % (model old)	% model 1 / expected good % (model old)	..

Как сравнить модели и найти лучшую

1. Как сравниваются модели?
2. Цена ошибки?

Давайте обсудим...

МегаФон