

Яндекс

Стань контрибьютором в Open Source

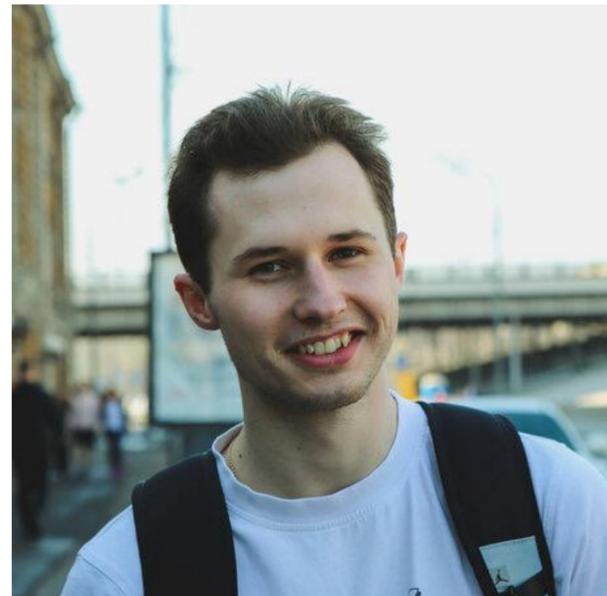
или как сделать свой первый pull request в CatBoost

Евгений Петров, ведущий разработчик

Команда CatBoost сегодня



Кирилл
Власов



Никита
Дмитриев



Екатерина
Ермишкина



Андрей
Хропов



Евгений
Петров

Как CatBoost учится

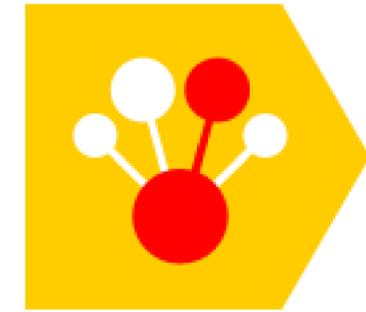
C++ в стиле Яндекс

Собираем CatBoost

CatBoost

The screenshot shows the GitHub repository page for `catboost / catboost`, which is public. At the top, there are buttons for `Edit Pins`, `Unwatch` (193), `Fork` (1.1k), and `Star` (7.2k). Below these are navigation tabs for `Code`, `Issues` (480), `Pull requests` (14), `Discussions`, `Actions`, and `Security`. The `Code` tab is active, showing a `master` branch selector, `Go to file`, `Add file`, and a highlighted `Code` button. A notification box on the left states: **Your master branch isn't protected**. It explains that the branch should be protected from force pushing or deletion, or require status checks before merging. It includes a `Learn more` link and buttons for `Protect this branch` and `Dismiss`. On the right, the `About` section describes the library as a fast, scalable, high performance Gradient Boosting on Decision Trees library, used for ranking, classification, regression and other machine learning tasks for Python, R, Java, C++. It also mentions support for computation on CPU and GPU.

Если коротко, то CatBoost

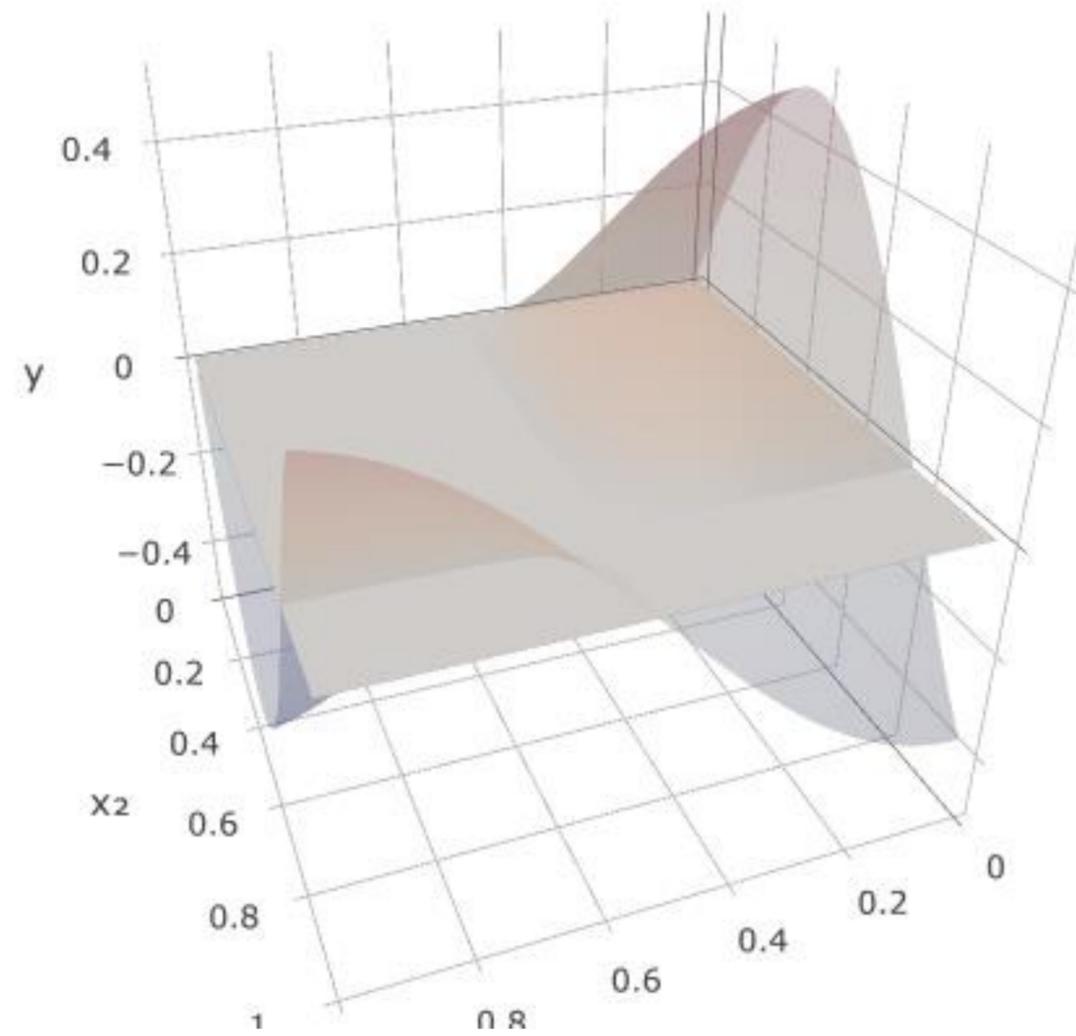


- 2017 г.р. проживает по адресу <https://github.com/catboost>
- Яндекс -- Поиск, Реклама, Алиса, Погода, Музыка, Беспилотники
- Open Source -- в Топ-200 из ~900К open source проектов по Google criticality score
- Машинное обучение -- Топ-3 на kaggle.com и ruri.org для табличных данных

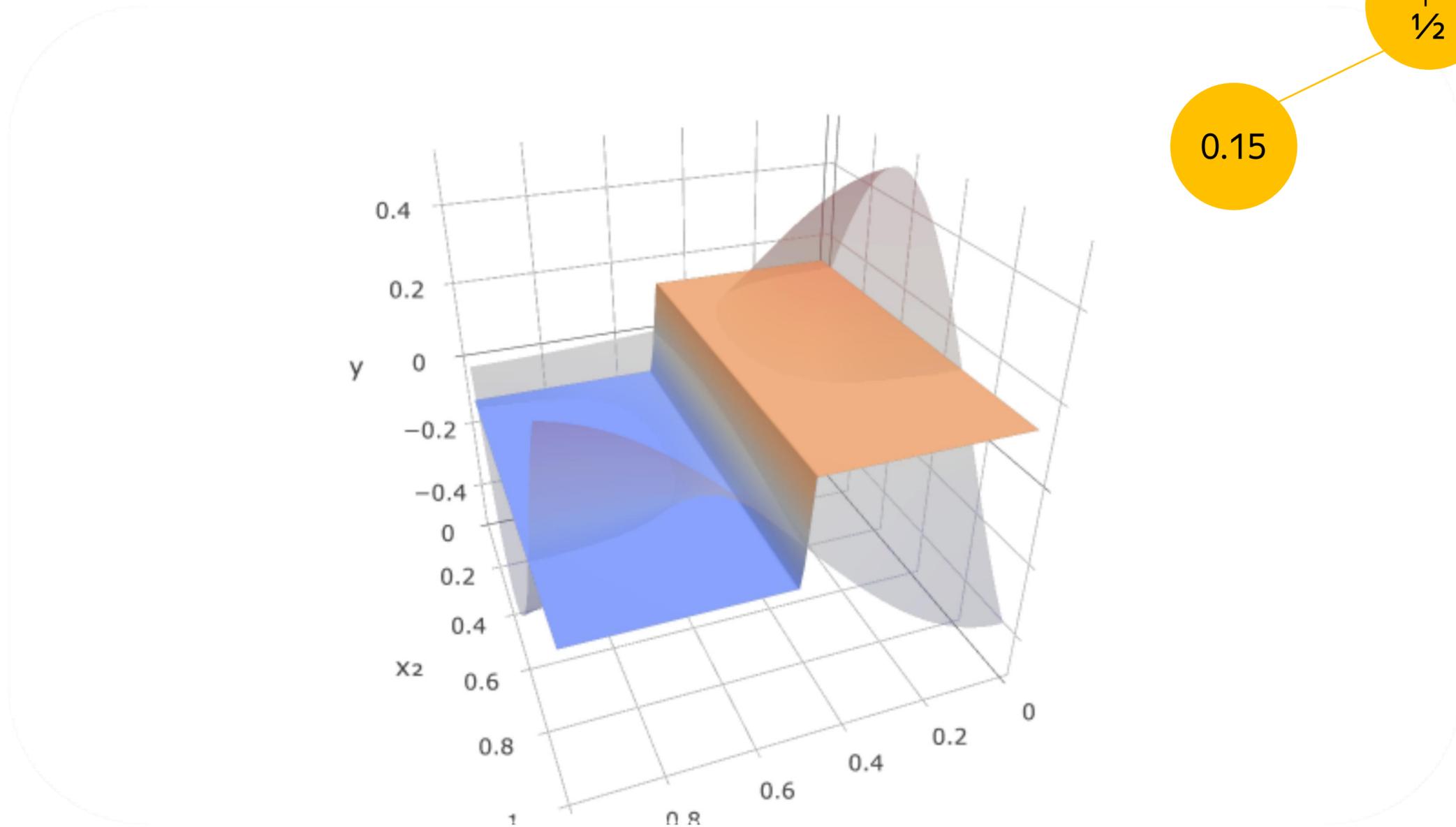
Как CatBoost учится



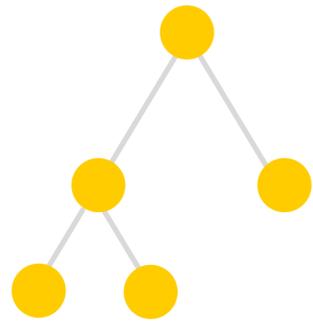
Градиентный бустинг на ... пальцах



Градиентный бустинг на ... пальцах

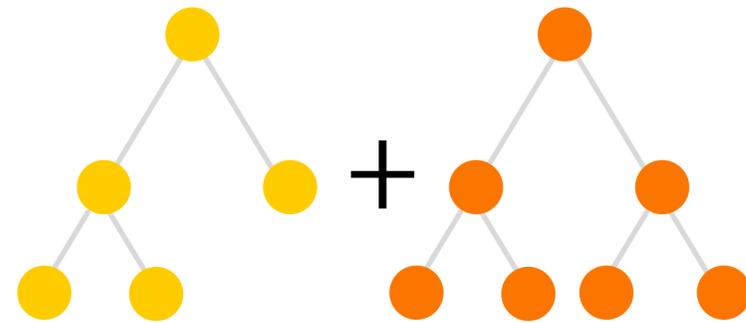
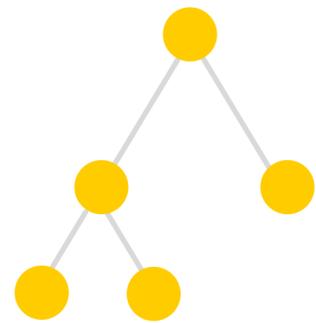


Градиентный бустинг на ... пальцах



Ошибка
предсказания
модели

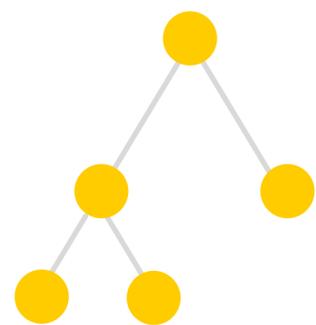
Градиентный бустинг на ... пальцах



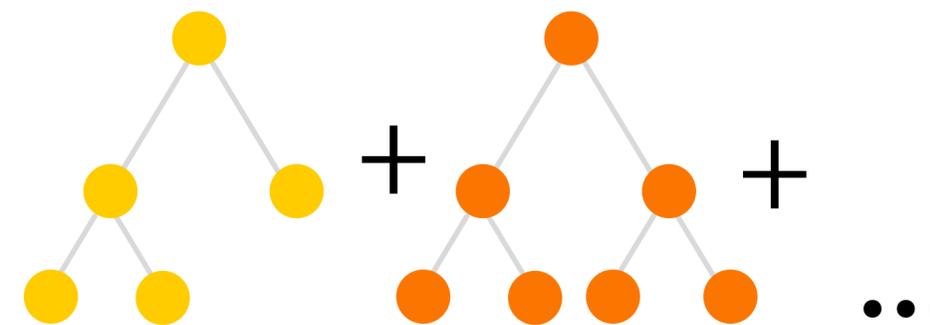
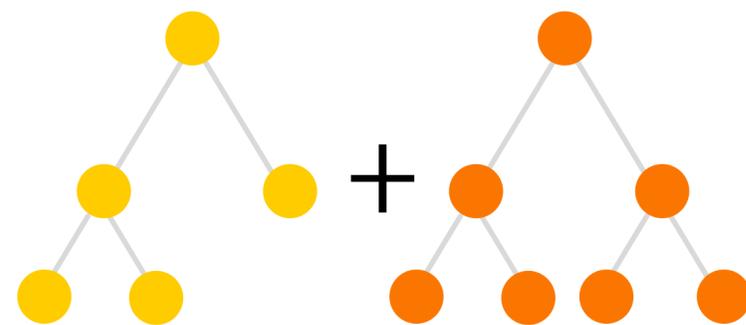
Ошибка
предсказания
модели



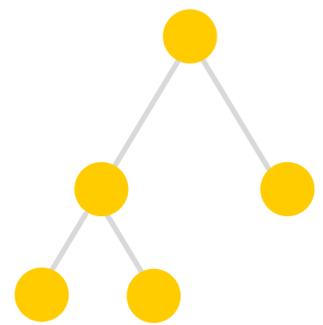
Градиентный бустинг на ... пальцах



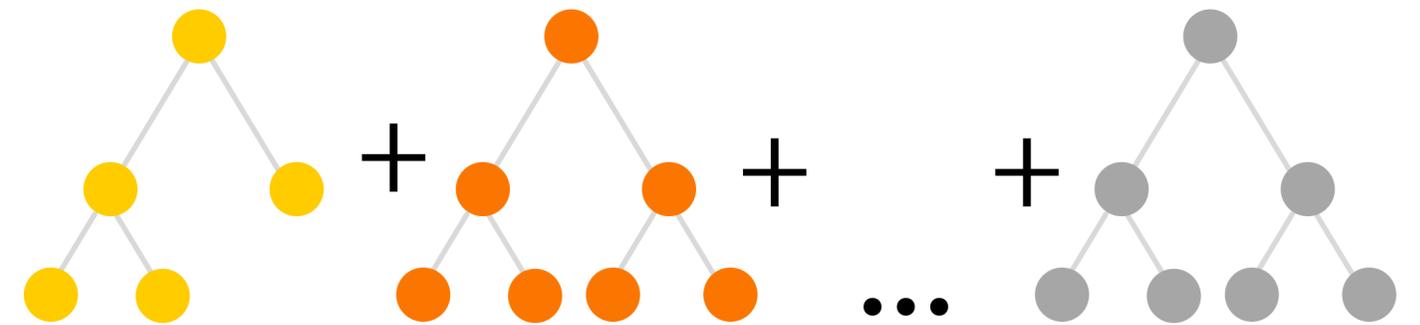
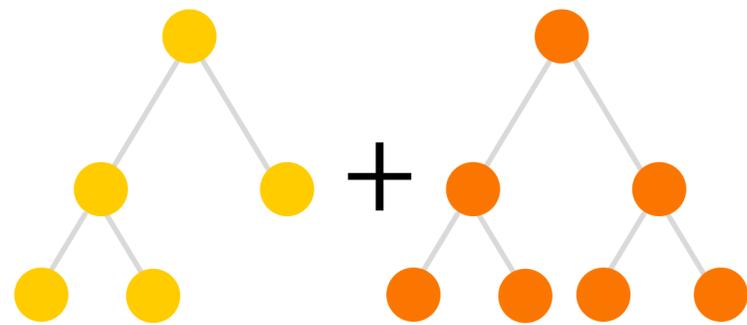
Ошибка
предсказания
модели



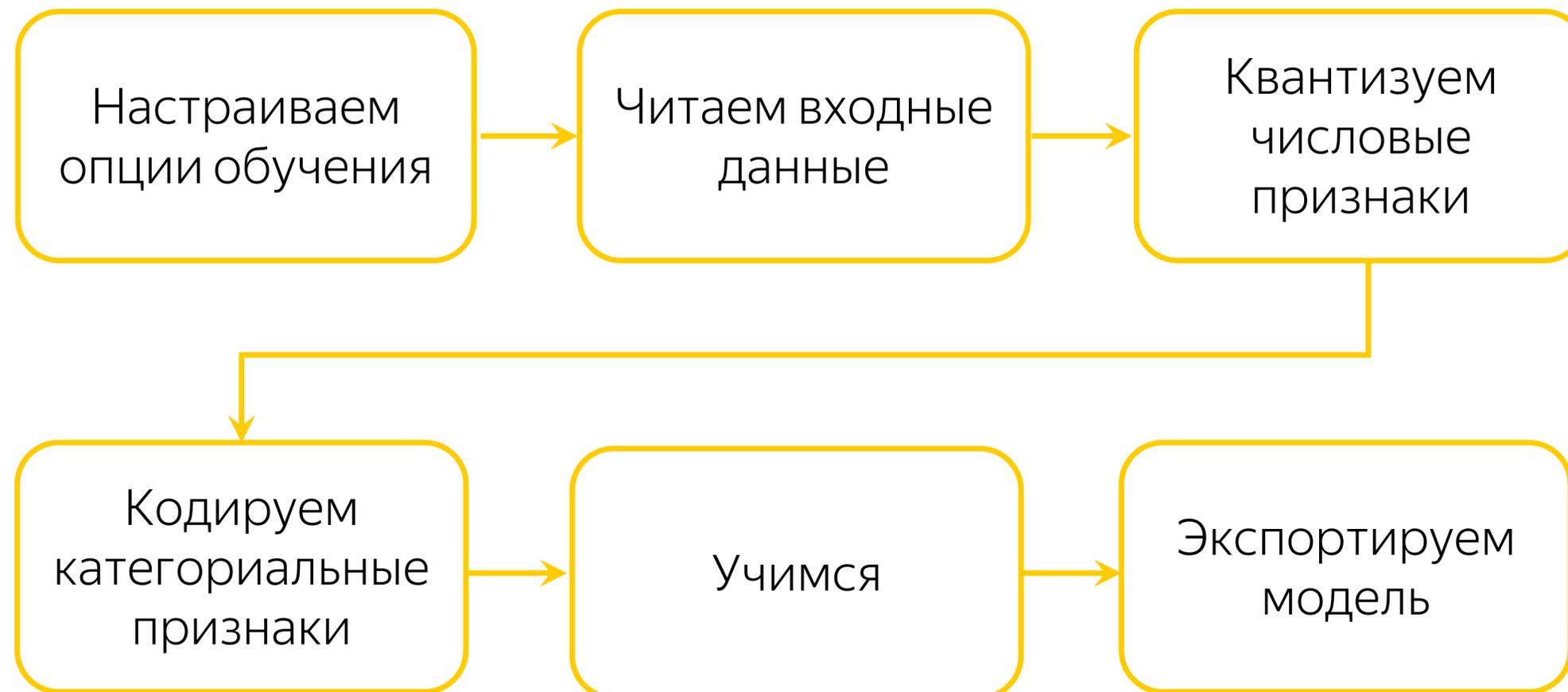
Градиентный бустинг на ... пальцах



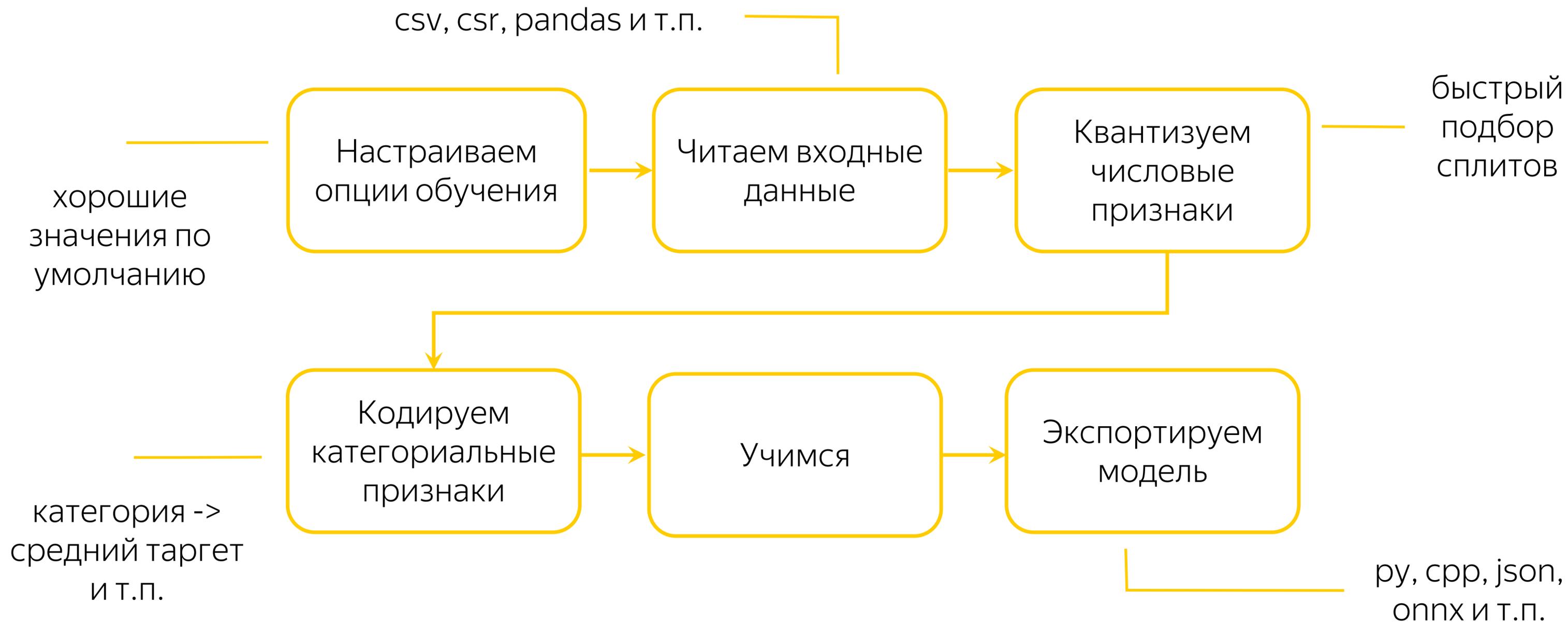
Ошибка
предсказания
модели



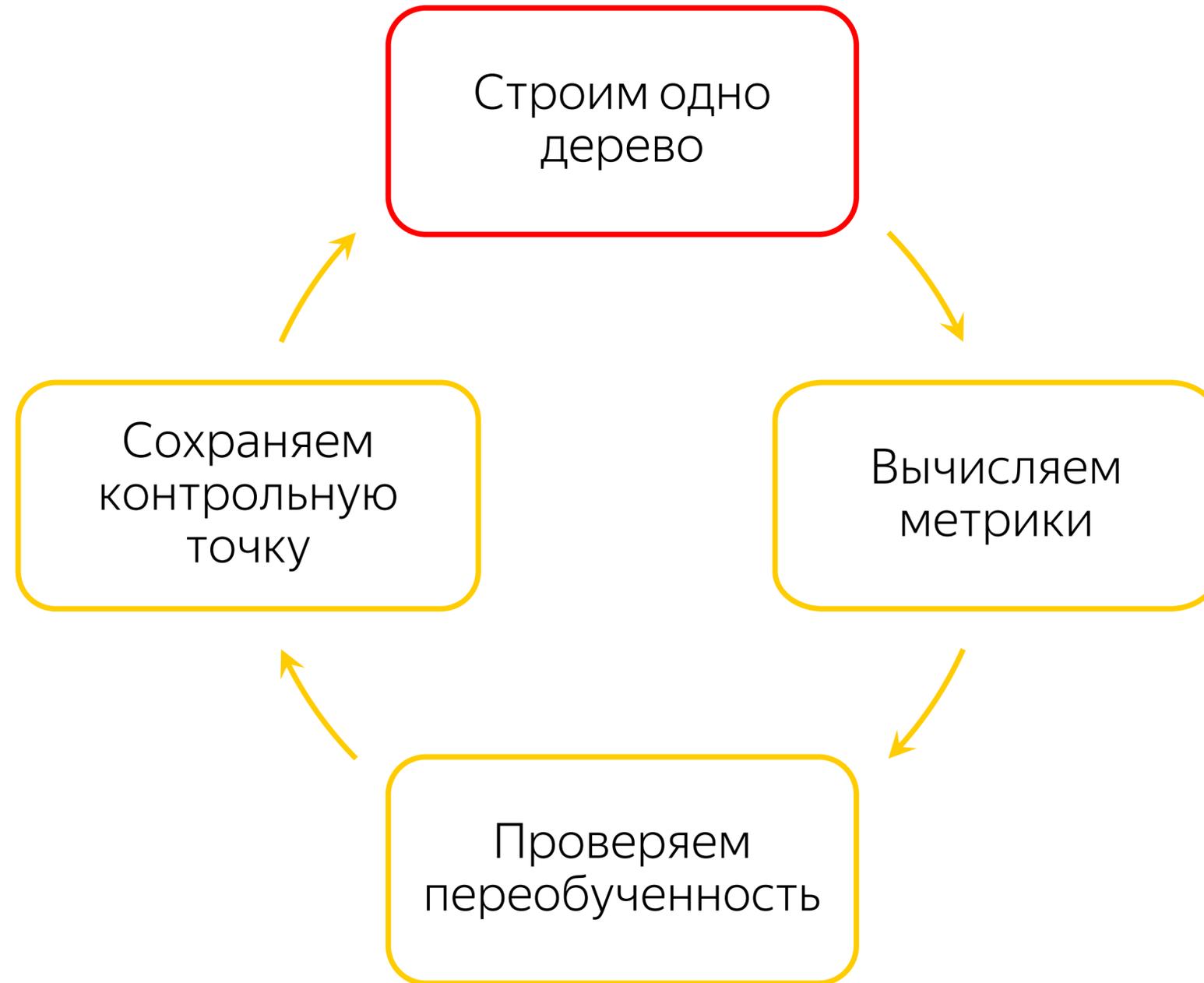
Процесс обучения



Процесс обучения



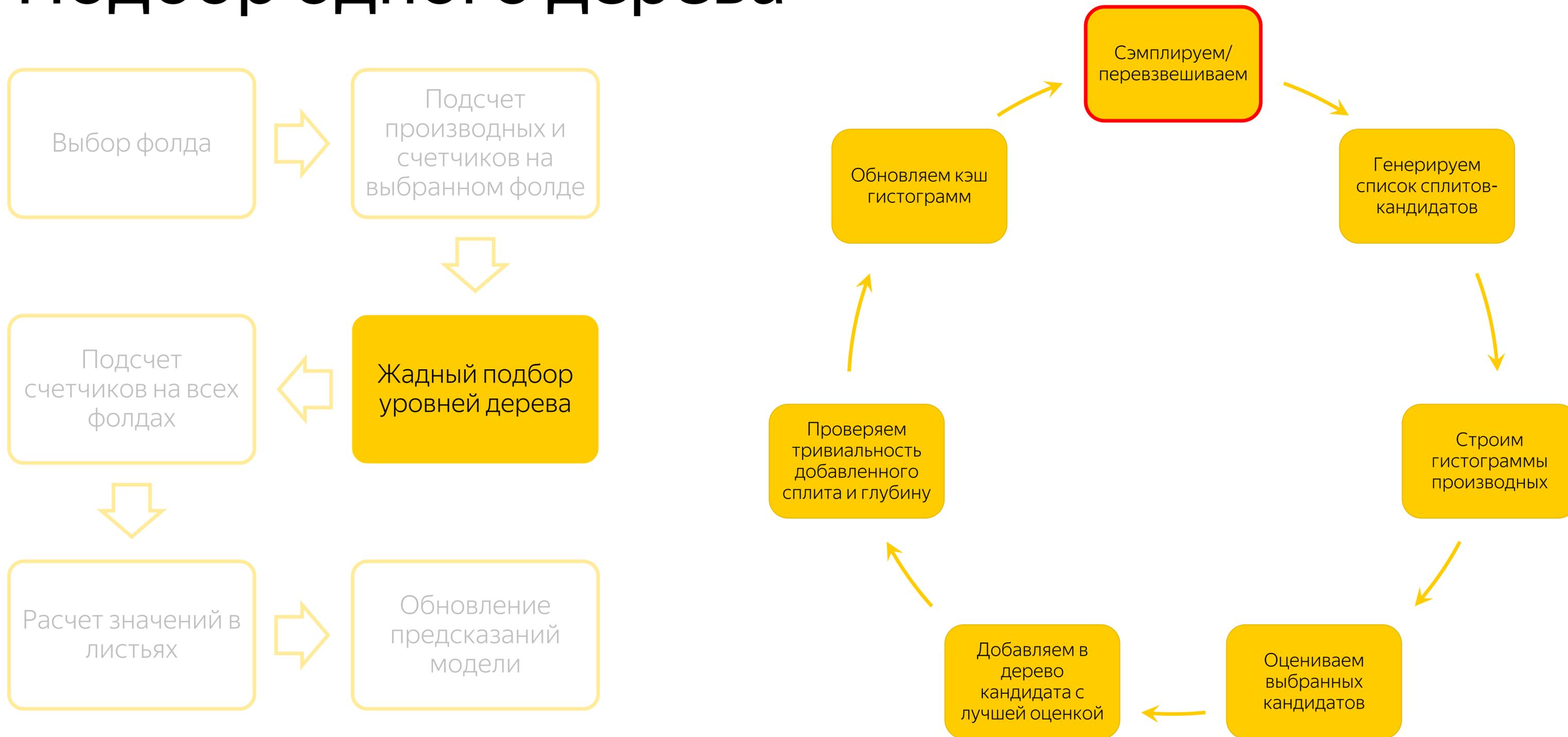
Главный цикл обучения



Подбор одного дерева



Подбор одного дерева



Структура <https://github.com/catboost/catboost>

catboost/

› CatBoost тут

catboost/docs/

› Дока по CatBoost тут

util/

› C++ контейнеры, файлы, строки, кодировки, ввод/вывод

library/

› общие библиотеки - json, protobuf, getopt, cuda libs, потоки

contrib/

› сторонние библиотеки - flatbuffers, onnx, coreml, nvidia cub

Важные места в коде (обучение на CPU)

catboost/libs/model/model.h
catboost/python_package/catboost/_catboost.pyx
catboost/python_package/catboost/core.py
catboost/R-package/src/catboost.cpp
catboost/private/libs/options

catboost/libs/train_lib/train_model.cpp
catboost/private/libs/algo/train.cpp
catboost/private/libs/algo/approx_calcer.cpp
catboost/private/libs/algo/score_calcer.cpp
catboost/private/libs/algo/greedy_tensor_search.cpp

Python -> Pyrex -> C++

Python -> Pyrex -> C++

```
def fit(self, X, y=None, cat_features=None, text_features=None,
        group_weight=None, subgroup_id=None, pairs_weight=None,
        eval_set=None, verbose=None, logging_level=None, plot=None,
        verbose_eval=None, metric_period=None, silent=None, early_stopping=None,
        save_snapshot=None, snapshot_file=None, snapshot_interval=None,
        log_cout=sys.stdout, log_cerr=sys.stderr):
    """
    Fit the CatBoost model.
```

catboost/python_package/catboost/core.py

Python -> Pyrex -> C++

```
def fit(self, X, y=None, cat_features=None, text_features=None,
        group_weight=None, subgroup_id=None, pairs_weight=None,
        eval_set=None, verbose=None, logging_level=None, plot=None,
        verbose_eval=None, metric_period=None, silent=None, early_stopping=None,
        save_snapshot=None, snapshot_file=None, snapshot_interval=None,
        log_cout=sys.stdout, log_cerr=sys.stderr):
```

```
"""
```

```
Fit the CatBoost model.
```

```
cpdef _train(self, _PoolBase train_pool, test_pools, dict params,
              self.model_blob = None
              _input_borders = params.pop("input_borders", None)
              prep_params = _PreprocessParams(params)
              cdef int thread_count = params.get("thread_count", 1)
              cdef TDataProviders dataProviders
              dataProviders.Learn = train_pool.__pool
              cdef _PoolBase test_pool
              cdef TVector[ui32] ignored_features
```

catboost/python_package/catboost/core.py

catboost/python_package/catboost/_catboost.pyx

Python -> Pyrex -> C++

```
def fit(self, X, y=None, cat_features=None, text_features=None,
        group_weight=None, subgroup_id=None, pairs_weight=None,
        eval_set=None, verbose=None, logging_level=None, plot=None,
        verbose_eval=None, metric_period=None, silent=None, early_stopping=None,
        save_snapshot=None, snapshot_file=None, snapshot_interval=None,
        log_cout=sys.stdout, log_cerr=sys.stderr):
    """
    Fit the CatBoost model.
```

catboost/python_package/catboost/core.py

catboost/python_package/catboost/_catboost.pyx

```
cpdef _train(self, _PoolBase train_pool, test_pools, dict params,
              self.model_blob = None
              _input_borders = params.pop("input_borders", None)
              prep_params = _PreprocessParams(params)
              cdef int thread_count = params.get("thread_count", 1)
              cdef TDataProviders dataProviders
              dataProviders.Learn = train_pool.Learn
              cdef _PoolBase test_pool
              cdef TVector[ui32] ignored
```

```
void TrainModel(
    NJson::TJsonValue plainJsonParams,
    NCB::TQuantizedFeaturesInfoPtr quantizedFeaturesInfo, // can
    const TMaybe<TCustomObjectiveDescriptor>& objectiveDescriptor,
    const TMaybe<TCustomMetricDescriptor>& evalMetricDescriptor,
    const TMaybe<TCustomCallbackDescriptor>& callbackDescriptor,
    NCB::TDataProviders pools, // not rvalue reference because Cy
    TMaybe<TFullModel*> initModel,
    THolder<TLearnProgress*> initLearnProgress,
```

catboost/libs/train_lib/train_model.cpp

Python -> Pyrex -> C++

```
def fit(self, X, y=None, cat_features=None, text_features=None,
        group_weight=None, subgroup_id=None, pairs_weight=None,
        eval_set=None, verbose=None, logging_level=None, plot=None,
        verbose_eval=None, metric_period=None, silent=None, early_stopping=None,
        save_snapshot=None, snapshot_file=None, snapshot_interval=None,
        log_cout=sys.stdout, log_cerr=sys.stderr):
    """
    Fit the CatBoost model.
```

catboost/python_package/catboost/core.py

```
cpdef _train(self, _PoolBase train_pool, test_pools, dict params,
              if.model_blob = None
              _input_borders = params.pop("input_borders", None)
              prep_params = _PreprocessParams(params)
              cdef int thread_count = params.get("thread_count", 1)
              cdef TDataProviders dataProviders
              dataProviders.Learn = train_pool.Learn
              cdef _PoolBase test_pool
              cdef TVector[ui32] ignored
```

catboost/python_package/catboost/_catboost.pyx

```
void TrainModel(
    NCB::TJsonValue plainJsonParams,
    NCB::TQuantizedFeaturesInfoPtr quantizedFeaturesInfo, // can
    const TMaybe<TCustomObjectiveDescriptor>& objectiveDescriptor,
    const TMaybe<TCustomMetricDescriptor>& evalMetricDescriptor,
    const TMaybe<TCustomCallbackDescriptor>& callbackDescriptor,
    NCB::TDataProviders pools, // not rvalue reference because Cy
    TMaybe<TFullModel*> initModel,
    THolder<TLearnProgress*> initLearnProgress,
```

catboost/libs/train_lib/train_model.cpp

C++ в стиле Яндекс



Почему в стиле Яндекс?

Почему в стиле Яндекс?

9 years ago r152	29	template <class T>
9 years ago r152	30	class TArrayRef: public NVector
9 years ago r152	31	public:
5 years ago r381	32	constexpr inline TArrayRef
9 years ago r152	33	: T_(nullptr)
	34	, s_(0)
	35	{
	36	}

std::span

Defined in header ``

```
template<
    class T,
    std::size_t Extent = std::dynamic_extent (since C++20)
> class span;
```

Почему в стиле Яндекс?

9 years ago r152	29	template <class T>
9 years ago r152	30	class TArrayRef: public NVector
9 years ago r152	31	public:
5 years ago r381	32	constexpr inline TArrayRef
9 years ago r152	33	: T_(nullptr)
	34	, s_(0)
	35	{
	36	}

std::span

Defined in header ``

```
template<
    class T,
    std::size_t Extent = std::dynamic_extent
> class span; (since C++20)
```

Примитивы C++ (умные указатели)

util/generic/ptr.h

- › `THolder<T>` – аналог `std::unique_ptr<T>`
- › `T(Atomic|Simple)SharedPtr<T>` – аналог `std::shared_ptr<T>`
- › `TIntrusivePtr<T>` – указатель на наследников класса `TRefCounted` + экономия на аллокации control-block

Примитивы C++ (ввод/вывод)

util/stream/*

- › `IInputStream` – базовый класс потоков ввода (`operator >>`)
- › `IOutputStream` – базовый класс потоков вывода (`operator <<`)
- › `Cin` – аналог `std::cin`
- › `Cout` – аналог `std::cout`
- › `Cerr` – аналог `std::cerr`
- › `endl` – аналог `std::endl`

Примитивы C++ (файлы)

util/stream/file.h

- › `TInputFile` – аналог `std::ifstream`
- › `TOutputFile` – аналог `std::ofstream`

util/system/fs.h

- › `NFs::Exists()` – проверка наличия файла/директории
- › `NFs::Copy()` – копировать файл

Примитивы C++ (контейнеры)

util/generic/vector.h

› TVector<T> – наследник `std::vector<T>`

util/generic/hash.h

› THashMap<T> – эквивалент `std::unordered_map<T>`

› THashSet<T> – эквивалент `std::unordered_set<T>`

util/generic/set.h + util/generic/map.h

› TSet<T> – наследник `std::set<T>`

› TMap<T> – наследник `std::map<T>`

Примитивы C++ (ссылки на массивы)

util/generic/array_ref.h

› `TArrayRef` и `TConstArrayRef` – альтернатива `std::span`
из C++20

Примитивы C++ (строки)

util/generic/strbuf.h

› TStringBuf – аналог std::string_view

util/generic/string.h

› TString – CoW строка char

› TUtf16String – CoW строка wchar16

util/string/cast.h

› TString ToString<T> – эквивалент std::to_string

› T FromString<T>

› bool TryFromString<T>(..., T* value)

Примитивы C++ (исключения и ассерты)

util/generic/yexception.h

- › `yexception` – наследник `std::exception`
- › `ythrow` – `throw` + печать стек трейса
- › `TString CurrentExceptionMessage()` – текстовое описание исключения

util/system/yassert.h

- › `Y_ASSERT()`
- › `Y_VERIFY()`

Примитивы C++ (сериализация)

util/ysaveload.h

- › Save(IOOutputStream*)
- › Load(IInputStream*)
- › автогенерация
Y_SAVELOAD(...)

Распределенное обучение на CPU

library/cpp/binsaver/bin_saver.h

- › T::operator&(IBinSaver*)
- › автогенерация SAVELOAD(...)

Примитивы C++ (безопасный union)

util/generic/maybe.h

› TMaybe – аналог std::optional

util/generic/variant.h

› TVariant – аналог std::variant

Исключения в CatBoost коде

catboost/libs/helpers/exception.h

- › `TCatBoostException` – `yexception` + стек трейс
- › `CB_ENSURE` – аналог `Y_ENSURE`, бросающий исключение `TCatBoostException`

Code style

Общий стиль кода на C++

- › https://github.com/catboost/catboost/blob/master/CPP_STYLE_GUIDE.md

Расширение стиля для catboost

- › Кроме util и catboost/cuda/*/kernel, разрешён C++17 и выше
- › https://github.com/catboost/catboost/blob/master/catboost_command_style_guide_extension.md

Python

- › PEP8 😊

Собираем CatBoost



Особенности сборки

Система сборки с версии CatBoost 1.2 – cmake

- › Стандартный прозрачный процесс сборки
- › Полностью отказались от `ya make`

Максимально статическая сборка

- › Минимум внешних зависимостей == простота использования артефактов сборки

Платформы сборки

Операционные системы

- › Linux, macOS, Windows, Android

Архитектуры

- › arm, x86
- › x86_64, aarch64 (arm64), ppc64le
- › x86_64-cuda, aarch64-cuda, ppc64le-cuda

Окружение для сборки

Окружение для сборки

Cmake 3.21

Conan 1.57.0 -- 1.59.0

Ninja

Окружение для сборки

Cmake 3.21

Conan 1.57.0 -- 1.59.0

Ninja

Linux

- › gcc, clang ≥ 12 , lld ≥ 7

macOS

- › XCode ≥ 12

Windows

- › Windows 10 или 11 SDK

- › MSVC v14.28

- › MSVS 2019 v16.8 или v16.9

Окружение для сборки

Cmake 3.21

Conan 1.57.0 -- 1.59.0

Ninja

CUDA 11.8, если нужно
обучение/применение на GPU

JDK 8.0, если нужна JVM часть

python-dev, если нужна Python
часть

Linux

- › gcc, clang \geq 12, lld \geq 7

macOS

- › XCode \geq 12

Windows

- › Windows 10 или 11 SDK

- › MSVC v14.28

- › MSVS 2019 v16.8 или v16.9

Окружение для сборки

Cmake 3.21

Conan 1.57.0 -- 1.59.0

Ninja

CUDA 11.8, если нужно
обучение/применение на GPU

JDK 8.0, если нужна JVM часть

python-dev, если нужна Python
часть

Linux

- › gcc, clang >= 12, lld >= 7

macOS

- › XCode >= 12

Windows

- › Windows 10 или 11 SDK

- › MSVC v14.28

- › MSVS 2019 v16.8 или v16.9

Примеры запуска сборки и тестов

Примеры запуска сборки и тестов

Приложение CatBoost без поддержки GPU

- › `python $CATBOOST_SRC_ROOT/build/build_native.py --build-root-dir=./build_no_cuda --targets catboost`

Примеры запуска сборки и тестов

Приложение CatBoost без поддержки GPU

- › `python $CATBOOST_SRC_ROOT/build/build_native.py --build-root-dir=./build_no_cuda --targets catboost`

Приложение CatBoost с поддержкой GPU

- › `python $CATBOOST_SRC_ROOT/build/build_native.py --build-root-dir=./build_with_cuda --targets catboost --have-cuda`

Примеры запуска сборки и тестов

Приложение CatBoost без поддержки GPU

- › `python $CATBOOST_SRC_ROOT/build/build_native.py --build-root-dir=./build_no_cuda --targets catboost`

Приложение CatBoost с поддержкой GPU

- › `python $CATBOOST_SRC_ROOT/build/build_native.py --build-root-dir=./build_with_cuda --targets catboost --have-cuda`

https://github.com/catboost/catboost/blob/master/build/build_native.py

Артефакты сборки

Расположение	Артефакт
catboost/app	обучение и применение из командной строки
catboost/python-package/catboost	динамическая библиотека для Питон-пакета
catboost/libs/model_interface	динамическая и статическая библиотека для применения моделей
catboost/libs/train_interface	динамическая библиотека для обучения моделей
catboost/jvm-packages/catboost4j-prediction/src/native_impl	JNI библиотека для применения
catboost/spark/catboost4j-spark/core/src/native_impl	JNI библиотека для обучения с поддержкой Apache Spark

Контрибьютировать в CatBoost – это просто

Контрибьютировать в CatBoost – это просто

- › Нужна учётка на github.com 😊

Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>

Контрибьютить в CatBoost – это просто

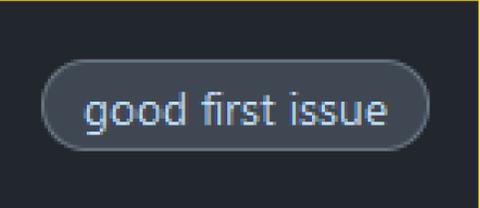
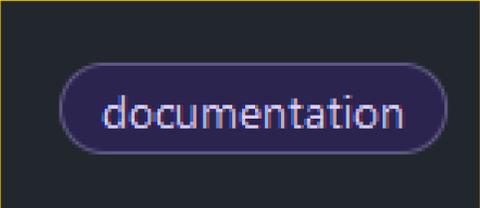
- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>

good first issue

documentation

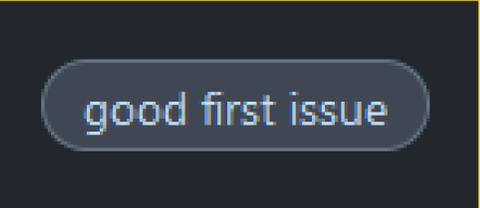
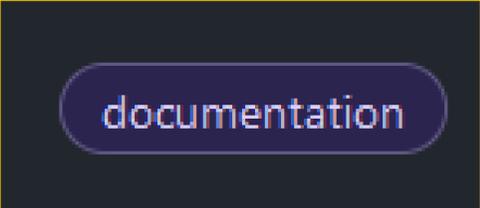
Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>
- › Если ещё нет, то делаем свой форк <https://github.com/catboost/catboost>

A dark blue rounded rectangular button with the text "good first issue" in a light blue font.A dark blue rounded rectangular button with the text "documentation" in a light blue font.

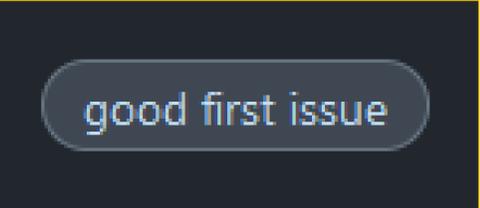
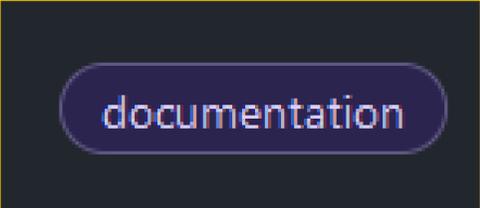
Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>
- › Если ещё нет, то делаем свой форк <https://github.com/catboost/catboost>
- › Делаем ветку для работы над задачей, заливаем в неё свои правки

A dark blue rounded rectangular button with the text "good first issue" in a light blue font.A dark blue rounded rectangular button with the text "documentation" in a light blue font.

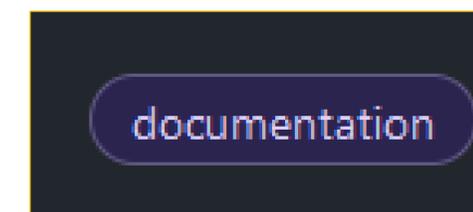
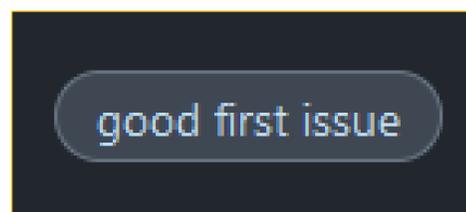
Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>
- › Если ещё нет, то делаем свой форк <https://github.com/catboost/catboost>
- › Делаем ветку для работы над задачей, заливаем в неё свои правки
- › Делаем pull request в основной репозиторий Катбуста

A dark blue rounded rectangular button with the text "good first issue" in a light blue font.A dark blue rounded rectangular button with the text "documentation" in a light blue font.

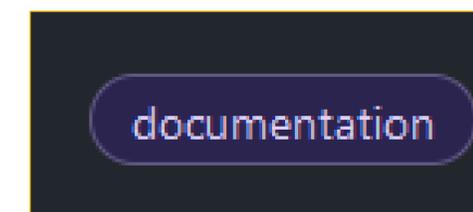
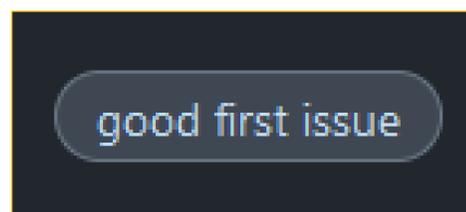
Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>
- › Если ещё нет, то делаем свой форк <https://github.com/catboost/catboost>
- › Делаем ветку для работы над задачей, заливаем в неё свои правки
- › Делаем pull request в основной репозиторий Катбуста
- › Исправляем замечания ревьюеров, в самый первый PR добавляем комментарий *I hereby agree to the terms of the CLA available at: [link]*.



Контрибьютить в CatBoost – это просто

- › Нужна учётка на github.com 😊
- › Выбираем понравившуюся задачу в <https://github.com/catboost/catboost/issues>
- › Если ещё нет, то делаем свой форк <https://github.com/catboost/catboost>
- › Делаем ветку для работы над задачей, заливаем в неё свои правки
- › Делаем pull request в основной репозиторий Катбуста
- › Исправляем замечания ревьюеров, в самый первый PR добавляем комментарий *I hereby agree to the terms of the CLA available at: [link]*.
- › Bingo!



Контрибьютить в CatBoost – это просто

CatBoost 1.2

- › Support CUDA applier in Rust package. #1925, thanks to @getumen
- › Add Focal loss (CPU-only for now). #1807, thanks to @diditforlulz273
- › Fix: model_interface/cmake_example failed build "runtime_error' is not a member of 'std". #2324, thanks to @Mandelag

CatBoost 1.0.5

- › Custom multilabel eval metrics, Fbeta score by @ELitvinova
- › Metrics plotter by @evgenabramov

CatBoost 1.0

- › Update C++ handles by reference to avoid redundant copies in R package by @david-cortes

Контакты

<https://catboost.ai/>

https://t.me/catboost_ru

https://t.me/catboost_en

<https://github.com/catboost/catboost>



Вопросы?

Евгений Петров

Ведущий разработчик

