

HW Production

Что сделано?

1. Трекер экспериментов

- 1.1. Развернут локально MLFlow с объектным хранилищем Minio и БД postgres ([Спасибо!](#))
- 1.2. Добавил метрики качества моделей по фолдам и залогировал параметры
- 1.3. Добавил для модели ALS обучение по эпохам, чтобы затреть loss на каждой итерации в mlflow
- 1.4. Затреть метрики по эпохам для Catboost
- 1.5. Залогировал время обучения и веса моделей
- 1.6. Выбрал лучшую модель для кандидатов из метрик, которые сохранялись в mlflow

2. Sentry

- 2.1. Подключил сервис к Sentry через их платформу
- 2.2. Эмулировал ошибки с user_id кратным 666 и моделью, которой нет в сервисе

3. Py-Spy

- 3.1. Снял три профиля сервиса: модель knn, popular и ручки /health
- 3.2. Проинтерпретировал результаты

4. ELK

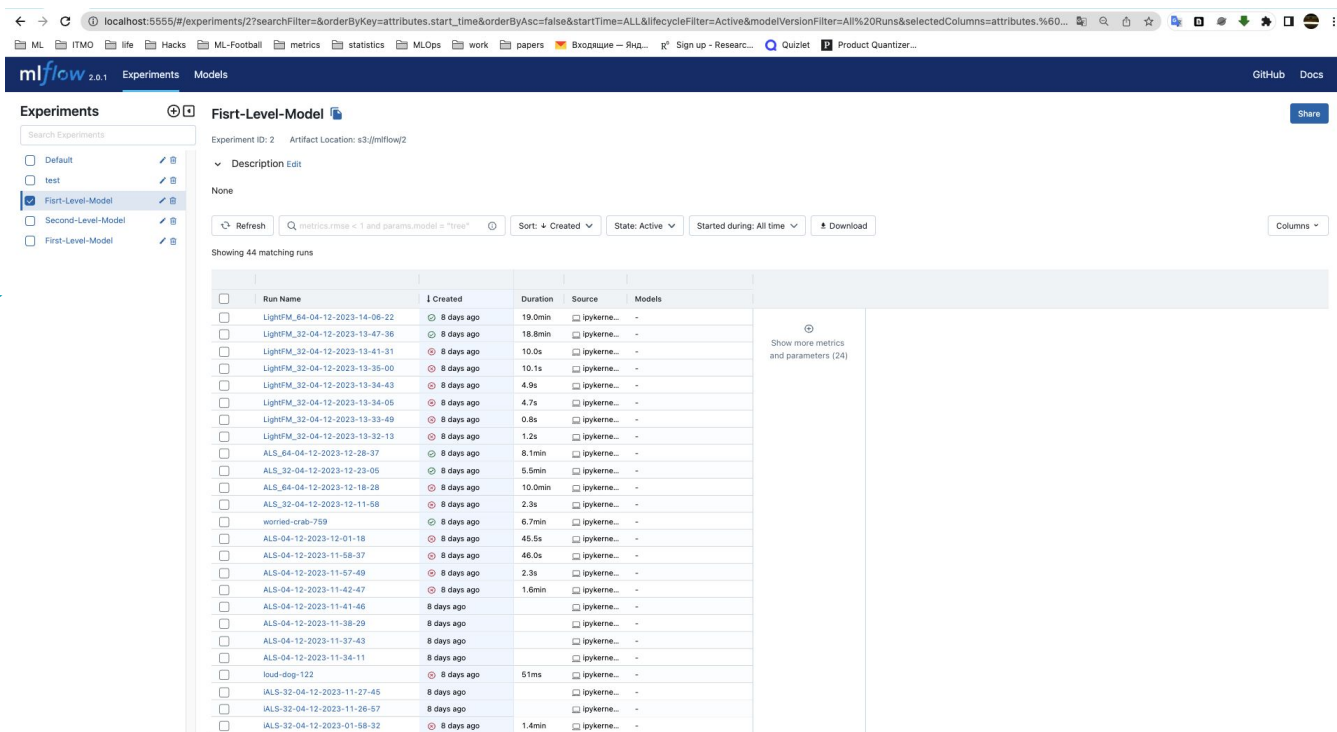
- 4.1. Обернул сервис в докер контейнер, и он в нём запустился
- 4.2. Развернул стек ELK с elasticsearch, kibana, filebeat без подключения к сервису ([Спасибо!](#))
- 4.3. Написал docker-compose ELK + сервис, но он не запускается :(

Трекинг экспериментов

Команда для запуска MLFlow - Minio - PostresSQL:

```
docker-compose up -d --build
```

Тестировал MLFlow



The screenshot displays the MLFlow web interface at localhost:5555. The 'Experiments' tab is active, showing a list of experiments on the left sidebar. The 'Fisrt-Level-Model' experiment is selected. The main panel shows the details of this experiment, including a description (none) and a list of 44 matching runs. The runs are sorted by 'Created' time, showing a list of runs with columns for Run Name, Created, Duration, Source, and Models. A 'Show more metrics and parameters (24)' link is visible on the right side of the runs table.

Run Name	Created	Duration	Source	Models
LightFM_64-04-12-2023-14-06-22	8 days ago	19.0min	ipykerne...	-
LightFM_32-04-12-2023-13-47-36	8 days ago	18.8min	ipykerne...	-
LightFM_32-04-12-2023-13-41-31	8 days ago	10.0s	ipykerne...	-
LightFM_32-04-12-2023-13-35-00	8 days ago	10.1s	ipykerne...	-
LightFM_32-04-12-2023-13-34-43	8 days ago	4.9s	ipykerne...	-
LightFM_32-04-12-2023-13-34-05	8 days ago	4.7s	ipykerne...	-
LightFM_32-04-12-2023-13-33-49	8 days ago	0.8s	ipykerne...	-
LightFM_32-04-12-2023-13-32-13	8 days ago	1.2s	ipykerne...	-
ALS_64-04-12-2023-12-28-37	8 days ago	8.1min	ipykerne...	-
ALS_32-04-12-2023-12-29-05	8 days ago	5.5min	ipykerne...	-
ALS_64-04-12-2023-12-18-28	8 days ago	10.0min	ipykerne...	-
ALS_32-04-12-2023-12-11-58	8 days ago	2.3s	ipykerne...	-
worried-crab-759	8 days ago	6.7min	ipykerne...	-
ALS-04-12-2023-12-01-18	8 days ago	45.5s	ipykerne...	-
ALS-04-12-2023-11-58-37	8 days ago	46.0s	ipykerne...	-
ALS-04-12-2023-11-57-49	8 days ago	2.3s	ipykerne...	-
ALS-04-12-2023-11-42-47	8 days ago	1.6min	ipykerne...	-
ALS-04-12-2023-11-41-46	8 days ago		ipykerne...	-
ALS-04-12-2023-11-38-29	8 days ago		ipykerne...	-
ALS-04-12-2023-11-37-43	8 days ago		ipykerne...	-
ALS-04-12-2023-11-34-11	8 days ago		ipykerne...	-
loud-dog-122	8 days ago	51ms	ipykerne...	-
ALS-32-04-12-2023-11-27-45	8 days ago		ipykerne...	-
ALS-32-04-12-2023-11-26-67	8 days ago		ipykerne...	-
ALS-32-04-12-2023-01-58-32	8 days ago	1.4min	ipykerne...	-

Трекинг экспериментов

Запуск для
моделей
кандидатов

localhost:5555/#/experiments/4?searchFilter=&orderByKey=attributes.start_time&orderByAsc=false&startTime=ALL&lifecycleFilter=Active&modelVersionFilter=All%20Runs&selectedColumns=attributes.%66...

mlflow 2.0.1 Experiments Models GitHub Docs

Experiments

Search Experiments

- ☐ Default
- ☐ test
- ☐ First-Level-Model
- ☐ Second-Level-Model
- ☒ First-Level-Model

First-Level-Model

Experiment ID: 4 Artifact Location: s3://mlflow/4

> Description Edit

Showing 4 matching runs

	Run Name	Created	User	Source	Version	Models	Metrics							
							MAP_at_10_by	NDCG_at_10_j	model_size_Mi	novelty_by_Fol	prec_at_10_by	recall_at_10_b	serendipity_by	time_fit
<input type="checkbox"/>	LightFM_64-04-15-2023-19-20-27	5 days ago	andrewsem...	ipykerne...	8d96af	-	0.063	0.033	487.1	5.742	0.03	0.167	5.722e-5	108.1
<input type="checkbox"/>	LightFM_32-04-15-2023-19-04-44	5 days ago	andrewsem...	ipykerne...	8d96af	-	0.072	0.037	247.3	5.437	0.032	0.174	5.416e-5	83.37
<input type="checkbox"/>	ALS_64-04-15-2023-18-50-46	5 days ago	andrewsem...	ipykerne...	8d96af	-	0.021	0.014	159.9	7.289	0.013	0.058	8.047e-5	95.13
<input type="checkbox"/>	ALS_32-04-15-2023-18-45-17	5 days ago	andrewsem...	ipykerne...	8d96af	-	0.02	0.014	79.94	6.733	0.013	0.06	5.503e-5	56.38

Трекинг экспериментов

Пример с ALS

← → ↺ localhost:5555/#/experiments/4/runs/aceb45e1fee84e7db05426d074c37dfd

ML ITMO life Hacks ML-Football metrics statistics MLOps work papers Входящие — Янд... R³ Sign up - Researc... Quizlet Product Quantizer...

Duration: 8.3min Status: FINISHED Lifecycle Stage: active

> Description [Edit](#)

▼ Parameters (2)

Name	Value
epochs	30
factors	64

▼ Metrics (13)

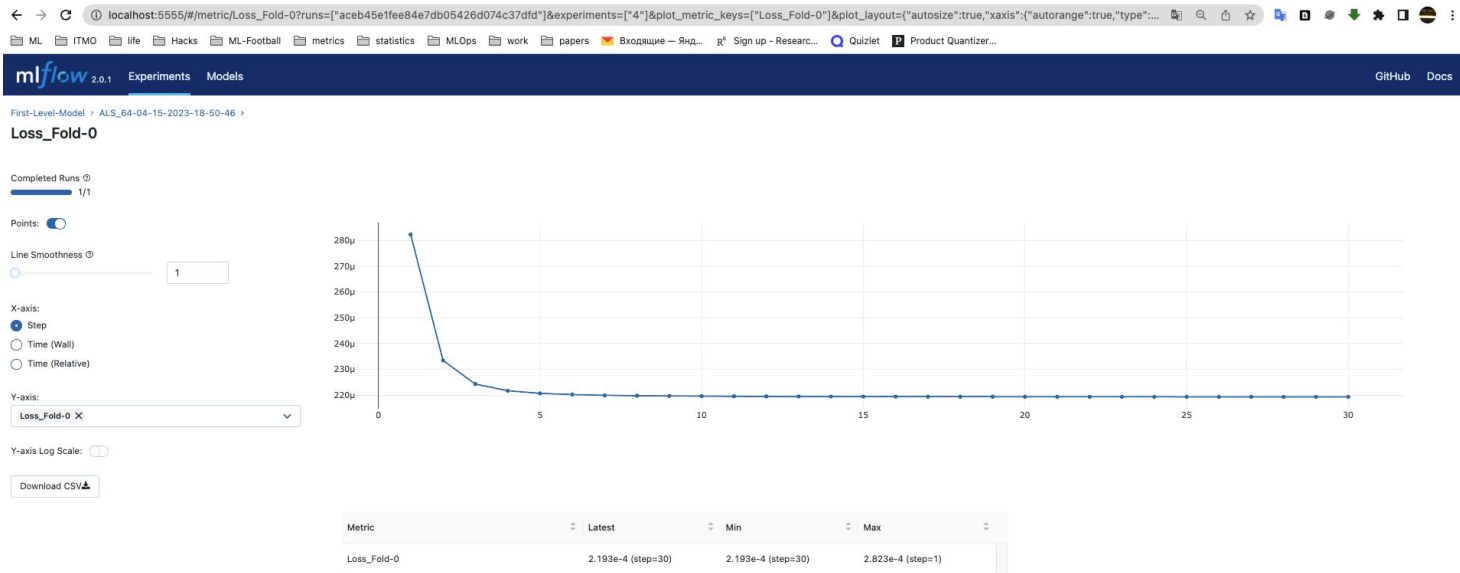
Name	Value
Loss_Fold-0 ↗	2.193e-4
Loss_Fold-1 ↗	2.132e-4
Loss_Fold-2 ↗	2.123e-4
Loss_Fold-3 ↗	2.078e-4
Loss_Fold-4 ↗	2.079e-4
MAP_at_10_by_Fold ↗	0.021
NDCG_at_10_by_Fold ↗	0.014
model_size_MB ↗	159.9
novelty_by_Fold ↗	7.289
prec_at_10_by_Fold ↗	0.013
recall_at_10_by_Fold ↗	0.058
serendipity_by_Fold ↗	8.047e-5
time_fit ↗	95.13

> Tags

▼ Artifacts

Трекинг экспериментов

Loss ALS по эпохам



Трекинг экспериментов

Модели для
ранжирования

The screenshot displays the mlflow Experiments interface. The top navigation bar includes 'Experiments' and 'Models'. The left sidebar shows a list of experiments: 'Default', 'test', 'First-Level-Model', 'Second-Level-Model' (selected), and 'First-Level-Model'. The main content area is titled 'Second-Level-Model' and shows 'Experiment ID: 3' and 'Artifact Location: s3://mlflow3'. Below this, there's a 'Description Edit' section. A search bar contains the query 'metrics.rmse < 1 and params.model = "tree"'. The interface includes filters for 'Sort: Created', 'State: Active', and 'Started during: All time'. A 'Download' button is also present. The table below shows 17 matching runs.

<input type="checkbox"/>	Run Name	Created	Duration	Source	Models
<input type="checkbox"/>	CBM_PointWise_all_data-04-16-2023-20-05-43	4 days ago	7.0min	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise_all_data-04-16-2023-20-04-52	4 days ago	47.4s	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise_all_data-04-16-2023-20-03-32	4 days ago	1.3min	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise_all_data-04-16-2023-19-57-53	4 days ago	5.6min	ipykerne...	-
<input checked="" type="checkbox"/>	CBM_PointWise-04-16-2023-01-19-32	4 days ago	6.1min	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise-04-16-2023-00-35-57	4 days ago	43.5min	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise-04-16-2023-00-26-32	4 days ago	9.4min	ipykerne...	-
<input type="checkbox"/>	CBM_PointWise-04-16-2023-00-25-50	4 days ago	49.2s	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-17-02	5 days ago		ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-15-13	5 days ago	1.7min	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-11-10	5 days ago	4.0min	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-10-43	5 days ago	26.6s	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-09-31	5 days ago	1.2min	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-08-49	5 days ago	40.5s	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-08-19	5 days ago	28.5s	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-14-06-16	5 days ago	2.0min	ipykerne...	-
<input type="checkbox"/>	Catboost-PointWise-04-15-2023-13-51-42	5 days ago	14.5min	ipykerne...	-

Showing 17 matching runs

Load more

Трекинг экспериментов

Пример одного из
запусков модели
ранжирования

The screenshot displays the mlflow web interface in a browser. The top navigation bar includes links for ML, ITMO, file, Hacks, ML-Football, metrics, statistics, ML Ops, work, papers, Brochure, Sign up, Research, Quiet, and Product Quantizer. The main header shows 'mlflow 1.2.1' and 'Experiments Models' with links to GitHub and Docs.

The experiment name is 'CBM_PointWise_all_data-04-16-2023-20-05-43'. Below it, the Run ID is '826a3189854347e98f27ac7d26cfc96', the Date is '2023-04-16 20:05:43', the Source is 'ipykernel_launcher.py', the Git Commit is '8d96af509382a647d214b9d15ae3c1c108d91cd', and the User is 'andrewsamenov'. The Duration is '7.0min' and the Status is 'FINISHED'. The Lifecycle Stage is 'active'.

There are two expandable sections: 'Parameters (9)' and 'Metrics (11)'. The 'Parameters' section shows a table with 9 rows, including 'custom_loss', 'learning_rate', 'max_depth', 'n_estimators', 'random_seed', 'subsample', 'thread_count', 'train_dir', and 'verbose'. The 'Metrics' section shows a table with 11 rows, including 'Epoch', 'Learn_Accuracy', 'Learn_Logloss', 'Learn_Precision', 'Learn_Recall', 'Test_AUC', 'Test_Accuracy', 'Test_Logloss', 'Test_Precision', 'Test_Recall', and 'Test_Results'.

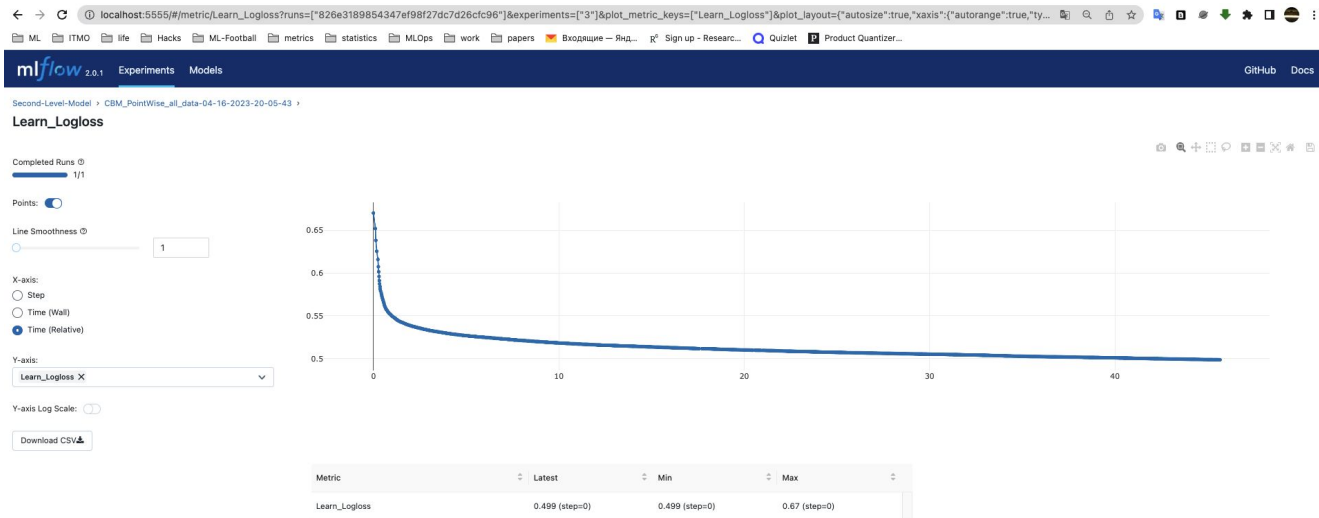
At the bottom, there are sections for 'Tags' and 'Artifacts'.

Name	Value
custom_loss	[AUC, 'Recall', 'Precision', 'Accuracy']
learning_rate	0.1
max_depth	4
n_estimators	3000
random_seed	17
subsample	0.9
thread_count	4
train_dir	./CBM_PointWise_all_data-04-16-2023-20-05-43/
verbose	100

Name	Value
Epoch	3000
Learn_Accuracy	0.758
Learn_Logloss	0.499
Learn_Precision	0.78
Learn_Recall	0.809
Test_AUC	0.82
Test_Accuracy	0.748
Test_Logloss	0.512
Test_Precision	0.773
Test_Recall	0.796
Test_Results	0.818

Трекинг экспериментов

Пример Loss во
время обучения
ранкера



Sentry

Эмулирование
отсутствия модели

The screenshot displays the Sentry web interface for an exception report. The browser address bar shows the URL: `itmo-1c.sentry.io/issues/4110592398/?query=is%3Aunresolved&referrer=issue-stream&stream_index=0`. The interface is divided into three main sections: a left sidebar, a central details pane, and a right sidebar.

Left Sidebar: Contains navigation links for ITMO (Andrey Semenov), Issues, Projects, Performance, Profiling, Replays, Crons, Alerts, Discover, Dashboards, Releases, User Feedback, Stats, and Settings. At the bottom, there is a 'Quick Start' section with 8 remaining tasks, and links for 'My Sentry Trial', 'Help', 'What's new', and 'Collapse'.

Central Details Pane: Displays the exception details for the issue. At the top, it shows the message 'Expected Exception object to report, got <class 'str'>!' with tabs for 'mechanism', 'logging', 'handled', and 'true'. Below this, the code snippet for `service/api/views.py` is shown, with line 78 highlighted: `capture_exception(f'Model name '{model_name}' not found')`. The exception type is `ModelNotFoundError` with the message `error_message='Model name '{model_name}' not found'`. The request details are shown in a JSON-like format: `{ 'asgi': { 'spec_version': '2.1', 'version': '3.0' }, 'client': [{ '127.0.0.1', 56125 },], 'http_version': '1.1', 'method': 'GET', 'path': '/reco/haha-model/176549', 'raw_path': 'b'/reco/haha-model/176549', 'root_path': '', 'scheme': 'http', 'server': [{ '127.0.0.1', 9994 },], 'type': 'http' }`. The `user_id` is 176549. At the bottom, it indicates the exception was called from `fastapi/routing.py` in `run_endpoint_function` and `service/api/middlewares.py` in `dispatch` at line 49.

Right Sidebar: Shows 'All Tags' with a list of tags and their values, all with a 100% occurrence rate. The tags include: `transaction` (generic FastAPI request), `generic FastAPI request`, `uri` (`http://localhost:9994/reco/haha-m...`), `release` (`a97ffa4d7644`), `browser` (Python Requests 2.25), `browser.name` (Python Requests), `environment` (production), `handled` (yes), `level` (error), `logger` (app), `mechanism` (logging), `runtime` (CPython 3.9.13), `runtime.name` (CPython), and `server_name` (MacBook-Pro-Andrej.local).

Sentry

Эмулирования
user_id кратного 666

itmo-1c.sentry.io/issues/4110665269/?query=is%3Aunresolved&referrer=issue-stream&stream_index=0

ML ITMO life Hacks ML-Football metrics statistics MLOps work papers Входщие — Янд... Sign up - Research... Quizlet Product Quantizer...

ITMO
Andrey Semenov

Issues
Projects
Performance
Profiling Beta
Replays new
Crons Beta
Alerts
Discover
Dashboards
Releases
User Feedback
Stats
Settings

Quick Start
6 Remaining tasks

My Sentry Trial
Help
What's new
Collapse

ValueError

Expected Exception object to report, got <class 'str'>!

mechanism logging handled true

service/api/views.py in get_reco at line 84 In App

```
79     raise ModelNotFoundError(  
80         error_message=f'Model name '{model_name}' not found'  
81     )  
82  
83     if user_id > 10 ** 6:  
84         capture_exception(f'User {user_id} not found')  
85  
86         raise UserNotFoundError(error_message=f'User {user_id} not found')  
87  
88         if user_id % 666 == 0:  
89             capture_exception(f'User {user_id} not found')  
90             raise UserNotFoundError(error_message=f'User {user_id} not found')
```

model_name 'bmp25'

request

```
{  
    asgi : {  
        spec_version : '2.1',  
        version : '3.0'  
    },  
    client : [  
        '127.0.0.1',  
        49626  
    ],  
    http_version : '1.1',  
    method : 'GET',  
    path : '/reco/bmp25/1002606',  
    raw_path : b'/reco/bmp25/1002606',  
    root_path : '',  
    scheme : 'http',  
    server : [  
        '127.0.0.1',  
        9989  
    ],  
}
```

transaction generic FastAPI request 100%

generic FastAPI request 100%

uri http://127.0.0.1:9989/reco/bmp25/1... 100%

release 45faa4571718 100%

browser Python Requests 2.25 100%

browser.name Python Requests 100%

environment production 100%

handled yes 100%

level error 100%

logger app 100%

mechanism logging 100%

runtime CPython 3.9.16 100%

runtime.name CPython 100%

server_name amsemenov-Vivobook-A... 100%

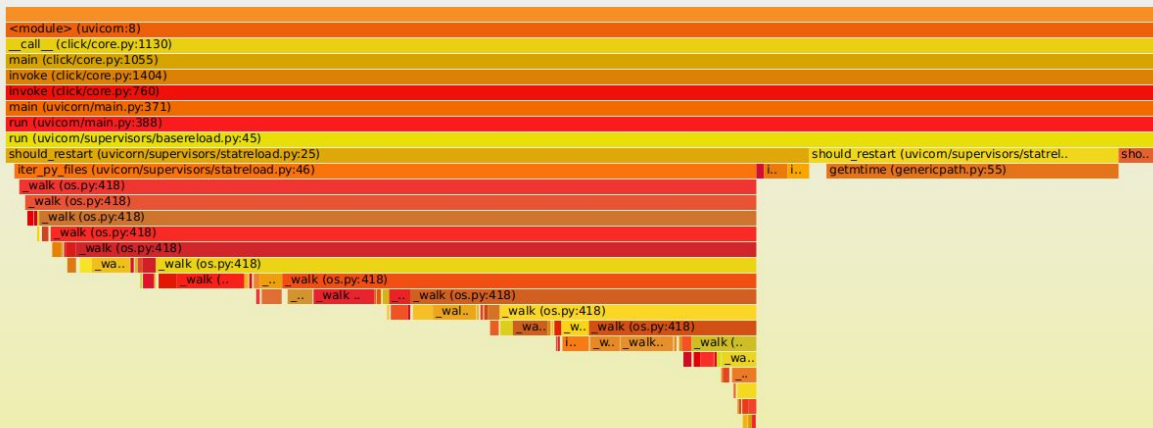
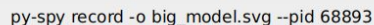
Py-Spy

Профиль для KNN

```
v/PycharmProjects/ITM0/RecoService2/.venv/bin/uvicorn main:app --reload --port 9994' (python v3.9.16)
Total Samples 22500
GIL: 6.00%, Active: 16.00%, Threads: 1
```

%Own	%Total	OwnTime	TotalTime	Function (filename)
10.00%	11.00%	21.83s	27.06s	_walk (os.py)
3.00%	3.00%	10.69s	10.69s	getmtime (genericpath.py)
1.00%	1.00%	3.94s	3.94s	islink (posixpath.py)
1.00%	16.00%	2.22s	42.15s	should_restart (uvicorn/supervisors/statreload.py)
1.00%	12.00%	2.16s	29.24s	iter_py_files (uvicorn/supervisors/statreload.py)
0.00%	0.00%	1.04s	1.29s	join (posixpath.py)
0.00%	0.00%	0.250s	0.250s	_get_sep (posixpath.py)
0.00%	16.00%	0.020s	42.18s	run (uvicorn/supervisors/basereload.py)
0.00%	0.00%	0.020s	0.020s	walk (os.py)
0.00%	0.00%	0.010s	0.010s	__enter__ (threading.py)
0.00%	16.00%	0.000s	42.18s	invoke (click/core.py)
0.00%	16.00%	0.000s	42.18s	<module> (uvicorn)
0.00%	0.00%	0.000s	0.010s	wait (threading.py)
0.00%	16.00%	0.000s	42.18s	__call__ (click/core.py)
0.00%	16.00%	0.000s	42.18s	main (uvicorn/main.py)
0.00%	16.00%	0.000s	42.18s	main (click/core.py)

```
Press Control-C to quit, or ? for help.
```



Интерпретация:

Много CPU тратится на процесс run в файле main и его последующих запусков частей сервиса, но не так значительно сри забирается uvicorn, потому что данная модель работает не так быстро

Py-Spy

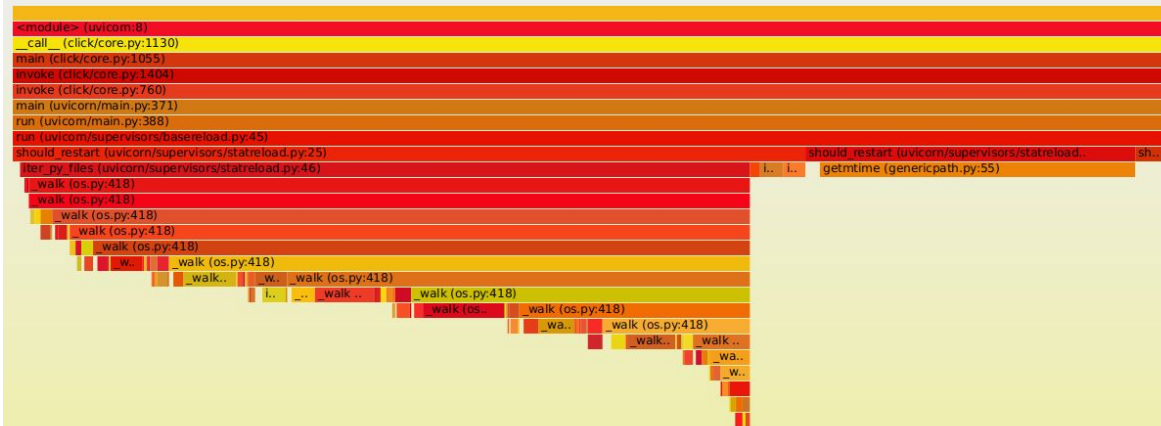
Профиль для
Popular

```
v/PycharmProjects/ITM0/RecoService2/.venv/bin/uvicorn main:app --reload --port 9993' (python v3.9.16)
Total Samples 10900
GIL: 8.00%, Active: 24.00%, Threads: 1
```

%Own	%Total	OwnTime	TotalTime	Function (filename)
8.00%	12.00%	10.26s	12.76s	_walk (os.py)
9.00%	9.00%	5.30s	5.30s	getmtime (genericpath.py)
3.00%	3.00%	2.01s	2.01s	islink (posixpath.py)
3.00%	24.00%	1.16s	20.21s	should_restart (uvicorn/supervisors/statreload.py)
0.00%	12.00%	0.990s	13.75s	iter_py_files (uvicorn/supervisors/statreload.py)
1.00%	1.00%	0.410s	0.490s	join (posixpath.py)
0.00%	0.00%	0.080s	0.080s	_get_sep (posixpath.py)
0.00%	0.00%	0.010s	0.010s	__exit__ (threading.py)
0.00%	24.00%	0.000s	20.22s	<module> (uvicorn)
0.00%	24.00%	0.000s	20.22s	main (click/core.py)
0.00%	24.00%	0.000s	20.22s	__call__ (click/core.py)
0.00%	24.00%	0.000s	20.22s	run (uvicorn/supervisors/basereload.py)
0.00%	24.00%	0.000s	20.22s	main (uvicorn/main.py)
0.00%	24.00%	0.000s	20.22s	invoke (click/core.py)
0.00%	24.00%	0.000s	20.22s	run (uvicorn/main.py)
0.00%	0.00%	0.000s	0.010s	wait (threading.py)

Press **Control-C** to quit, or **?** for help.

py-spy record -o popular.svg --pid 69100



Интерпретация:

Много CPU тратится на процесс run в файле main и сам uvicorn, т.к. рекомендации отдаются очень быстро, не давая отдельным компонентам сервиса передохнуть

Py-Spy

Профиль для ручки
/health

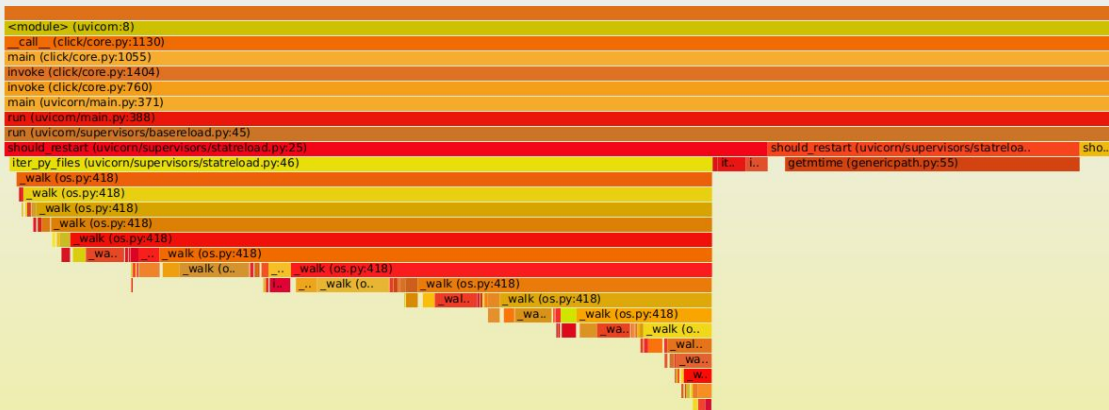
```
v/PycharmProjects/ITMO/RecoService2/.venv/bin/uvicorn main:app --reload --port 9992' (python v3.9.16)  
Total Samples 4800  
GIL: 6.00%, Active: 21.00%, Threads: 1
```

%Own	%Total	OwnTime	TotalTime	Function (filename)
8.00%	13.00%	4.52s	5.62s	_walk (os.py)
5.00%	5.00%	2.24s	2.24s	getmtime (genericpath.py)
3.00%	3.00%	0.830s	0.830s	islink (posixpath.py)
2.00%	21.00%	0.450s	8.72s	should_restart (uvicorn/supervisors/stateload.py)
1.00%	14.00%	0.400s	6.03s	iter_py_files (uvicorn/supervisors/stateload.py)
2.00%	2.00%	0.260s	0.270s	join (posixpath.py)
0.00%	0.00%	0.010s	0.010s	_get_sep (posixpath.py)
0.00%	0.00%	0.010s	0.010s	walk (os.py)
0.00%	21.00%	0.000s	8.72s	main (click/core.py)
0.00%	21.00%	0.000s	8.72s	__call__ (click/core.py)
0.00%	21.00%	0.000s	8.72s	run (uvicorn/main.py)
0.00%	21.00%	0.000s	8.72s	run (uvicorn/supervisors/basereload.py)
0.00%	21.00%	0.000s	8.72s	<module> (uvicorn)
0.00%	21.00%	0.000s	8.72s	invoke (click/core.py)
0.00%	21.00%	0.000s	8.72s	main (uvicorn/main.py)

Press **Control-C** to quit, or **?** for help.

py-spy record -o profile.svg --pid 68303

Search



Интерпретация:

Также много CPU тратится на процесс run в файле main, но сам uvicorn мало забирает ресурсов, т.к. нет обращения к выдаче рекомендаций, что позволяет моделям не работать