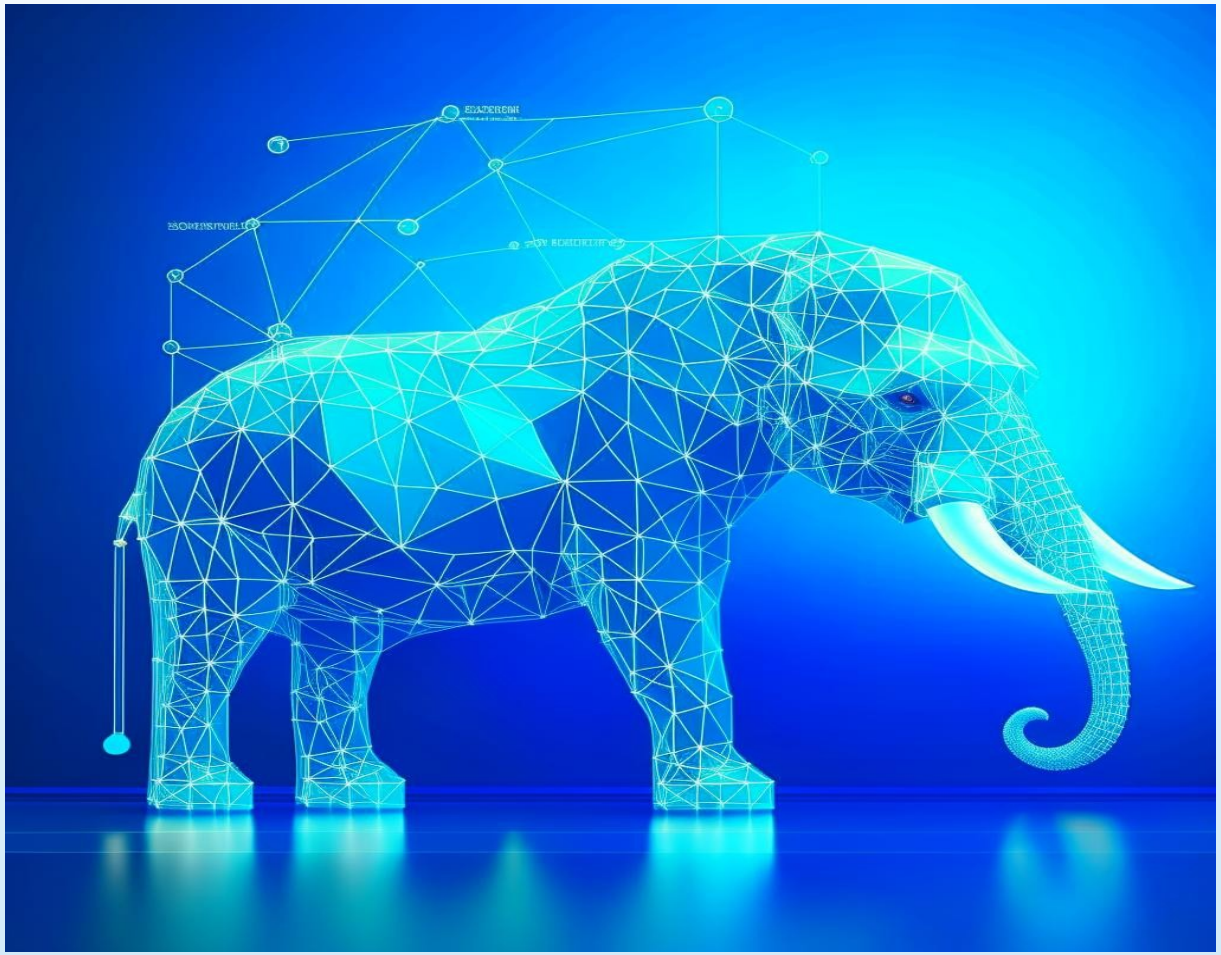
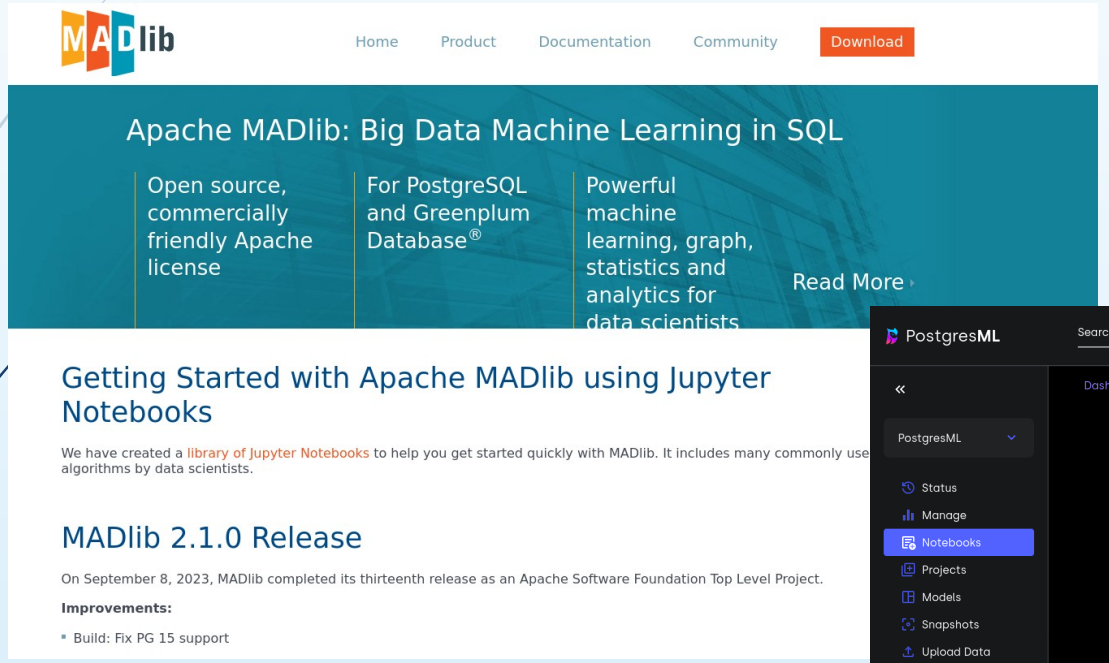
pg_ml

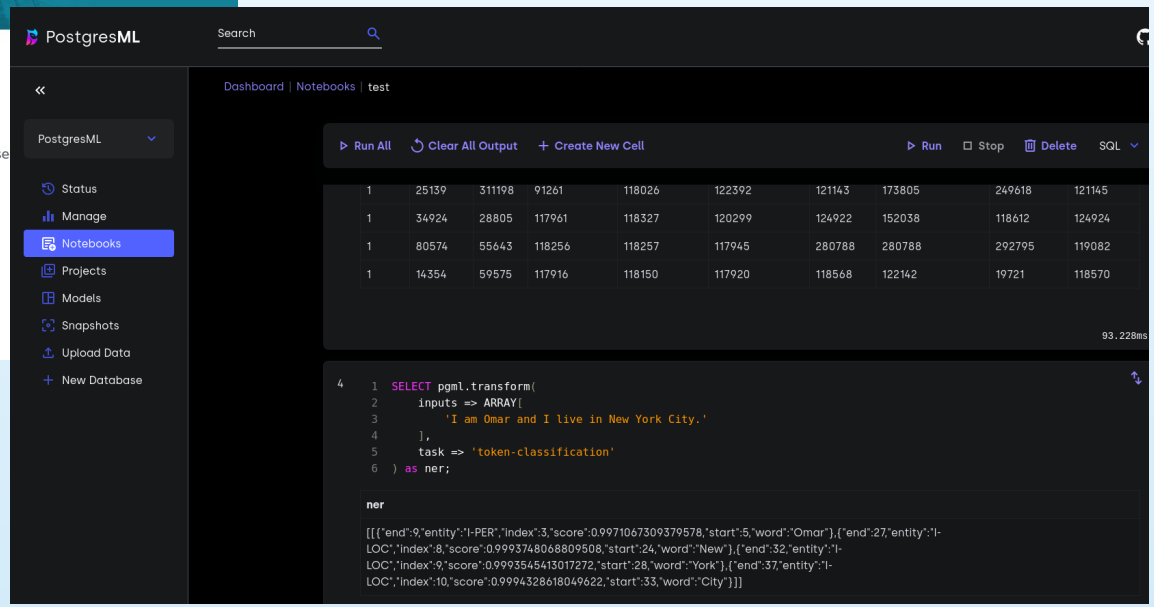


Почему БД?

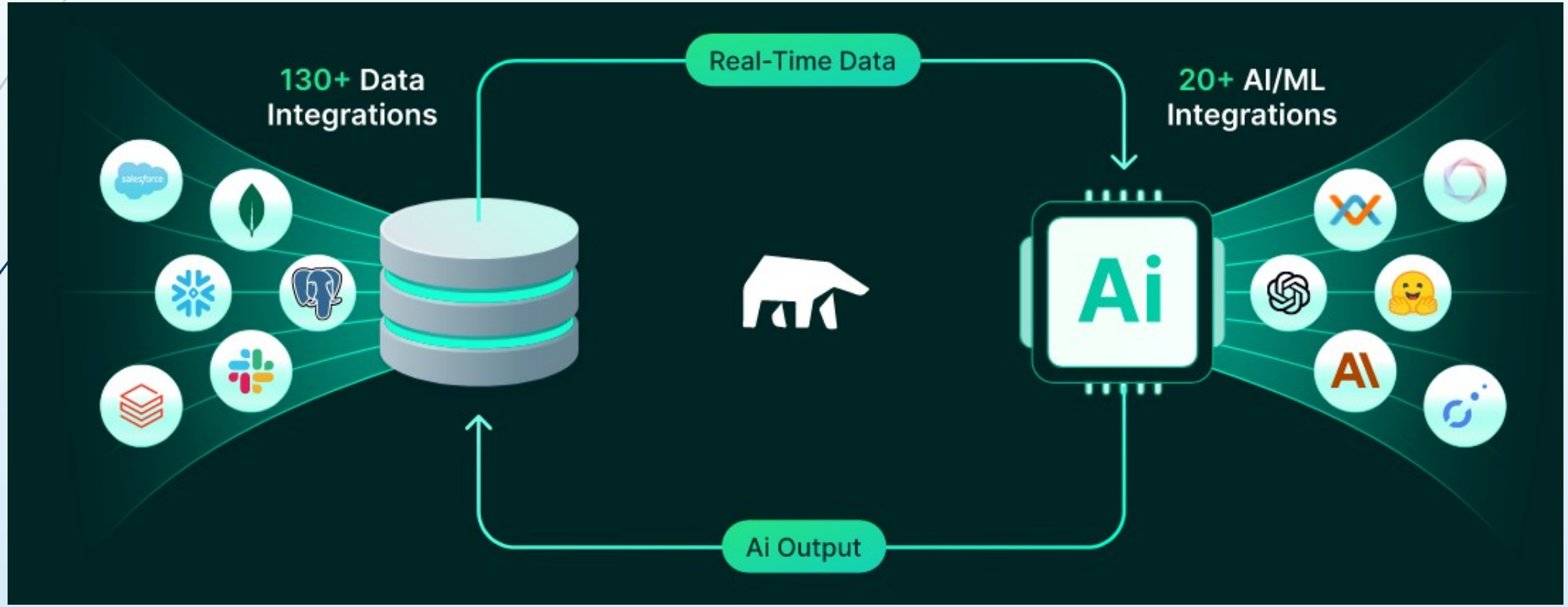
- Табличные данные
- Отсутствие промежуточных скриптов
- Все результаты получаются в запросах

Аналоги

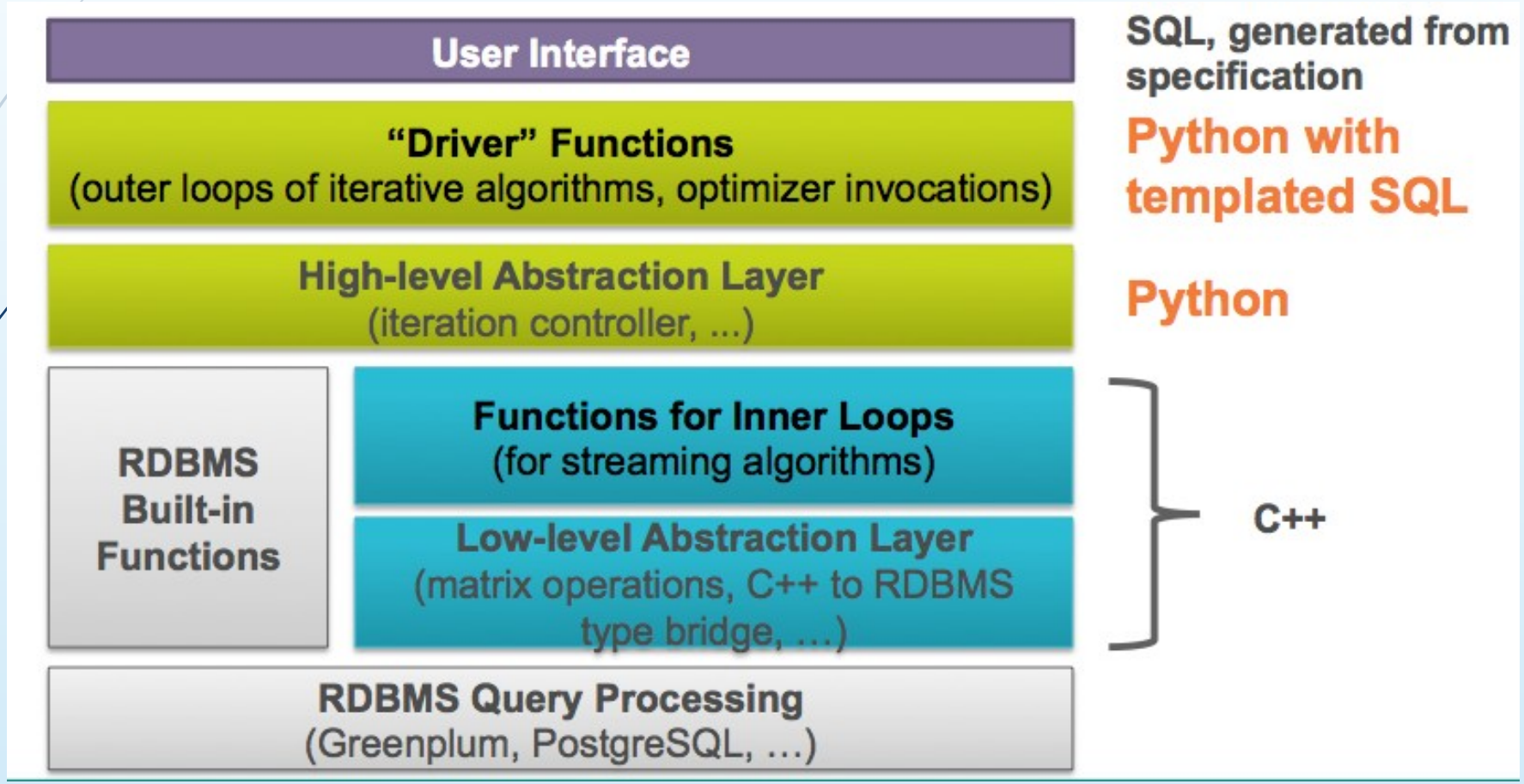

MADlib
Home
Product
Documentation
Community
Download
Apache MADlib: Big Data Machine Learning in SQL
Open source, commercially friendly Apache license
For PostgreSQL and Greenplum Database®
Powerful machine learning, graph, statistics and analytics for data scientists
Read More
Getting Started with Apache MADlib using Jupyter Notebooks
We have created a library of Jupyter Notebooks to help you get started quickly with MADlib. It includes many commonly use
algorithms by data scientists.
MADlib 2.1.0 Release
On September 8, 2023, MADlib completed its thirteenth release as an Apache Software Foundation Top Level Project.
Improvements:
Build: Fix PG 15 support


PostgresML
Search
Dashboard | Notebooks | test
PostgresML
Status
Manage
Notebooks
Projects
Models
Snapshots
Upload Data
New Database
Run All
Clear All Output
Create New Cell
Run
Stop
Delete
SQL
1 25139 311198 91261 118026 122392 121143 173805 249618 121145
1 34924 28805 117961 118327 120299 124922 152038 118612 124924
1 80574 55643 118256 118257 117945 280788 280788 292795 119082
1 14354 59575 117916 118150 117920 118568 122142 19721 118570
93.228ms
SELECT pgml.transform(
inputs => ARRAY[
'I am Omar and I live in New York City.'
],
task => 'token-classification'
) as ner;
ner
[[{"end":9,"entity":"I-PER","index":3,"score":0.9971067309379578,"start":5,"word":"Omar"},{"end":27,"entity":"I-LOC","index":8,"score":0.9993748068809508,"start":24,"word":"New"},{"end":32,"entity":"I-LOC","index":9,"score":0.9993545413017272,"start":28,"word":"York"},{"end":37,"entity":"I-LOC","index":10,"score":0.9994328618049622,"start":33,"word":"City"}]]

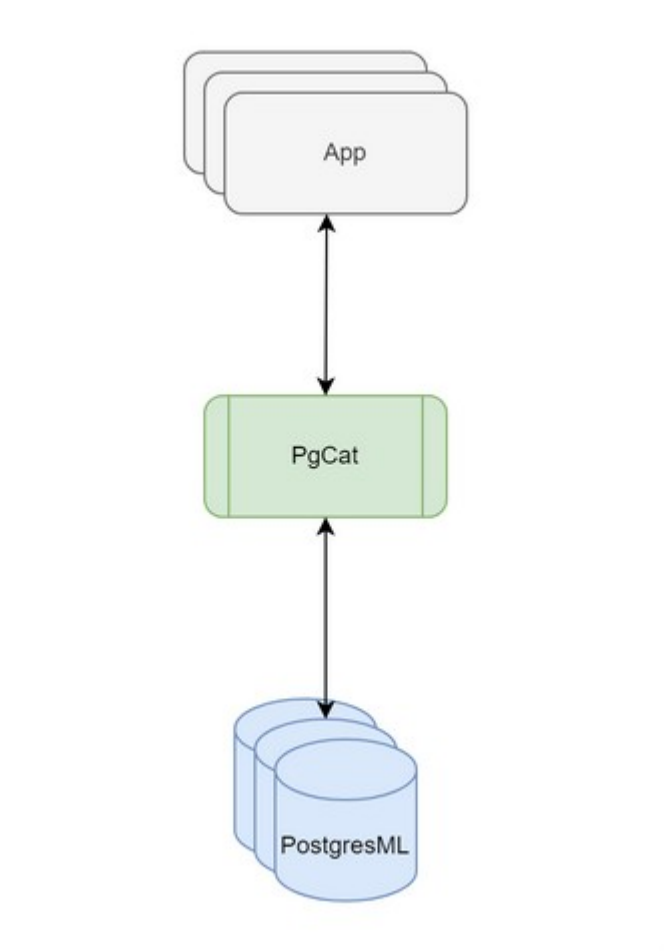
Аналоги


130+ Data Integrations
Real-Time Data
20+ AI/ML Integrations
Ai
Ai Output

MadLib


User Interface
"Driver" Functions
(outer loops of iterative algorithms, optimizer invocations)
High-level Abstraction Layer
(iteration controller, ...)
RDBMS Built-in Functions
Functions for Inner Loops
(for streaming algorithms)
Low-level Abstraction Layer
(matrix operations, C++ to RDBMS type bridge, ...)
RDBMS Query Processing
(Greenplum, PostgreSQL, ...)
SQL, generated from specification
Python with templated SQL
Python
C++

PostgresML

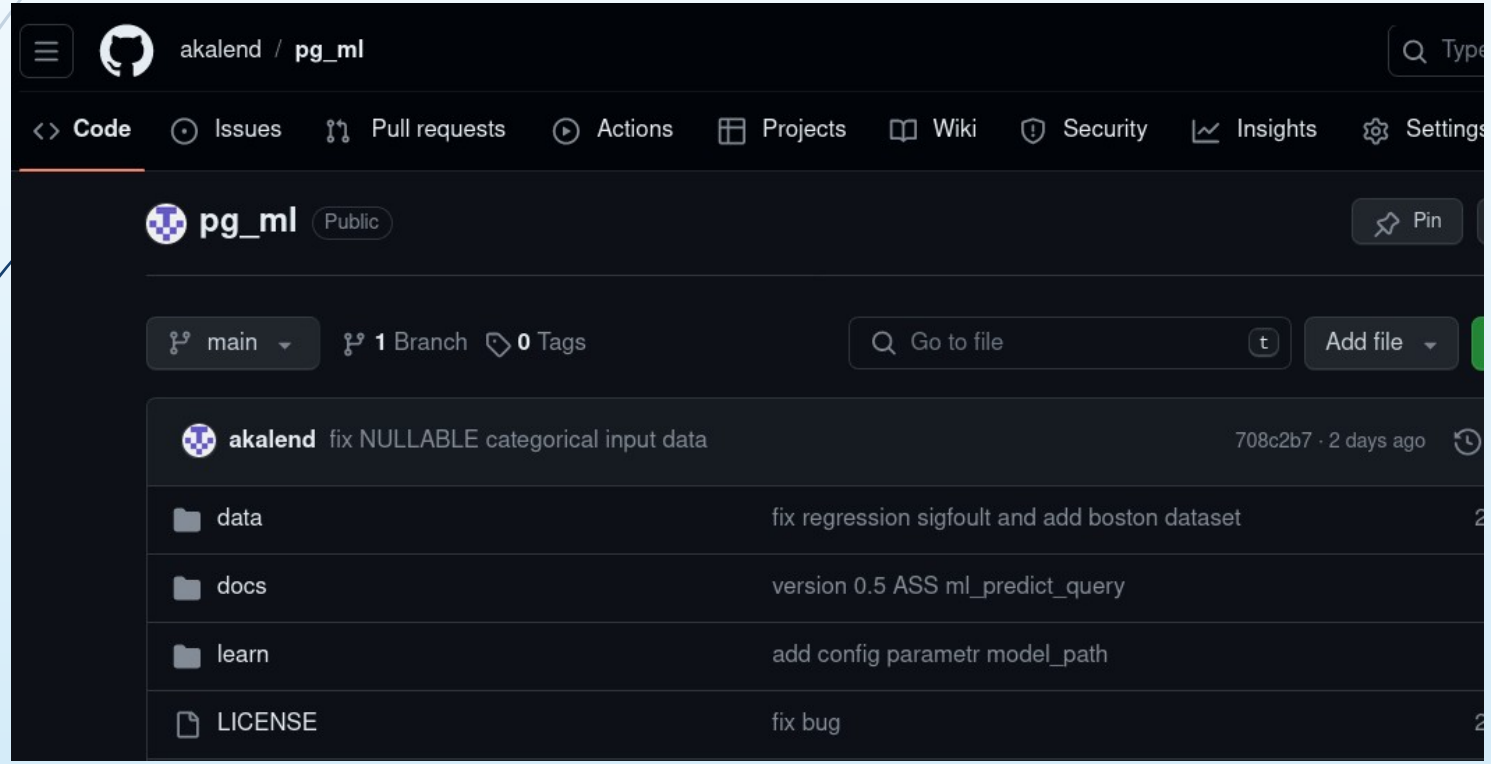


Недостатки Аналогов

- Использование API Python
- Не дружественный интерфейс работы с категорийными параметрами
- Нет множественной регрессии

О проекте:

Полностью открытый код



Доступны модели:

- Binary Classification
- Multi Classification
- Regression
- Ranking (идут работы)
- Text classification (идут работы)

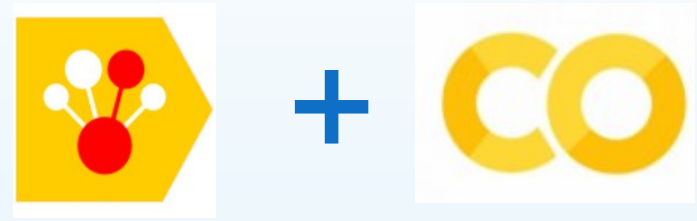
Основан на разработках Яндекс

CatBoost: https://catboost.ai/
libcatboostmodel.so

реализовано только предсказание
(ограничение libcatboostmodel)

ML процесс

1. Тенировка модели
2. Сохранение на сервер БД
3. предсказание



```
model.save_model('astra.cbm')
!ls
```

```
astra.cbm  astra.json  catboost_i
```

```
adult=# SELECT ml_predict ('astra3.cbm','astra3');
WARNING:  field run_id not used
WARNING:  field field_id not used
WARNING:  field spec_obj_id not used
WARNING:  field predict not used
      ml_predict
-----------------------
 public.astra3_predict
(1 row)
```

Соответсвия полей модели и таблицы в БД

ID, Id, id

FieldID, Field_ID

Import-port

Port, PORT

<=>

id

field_id

import_port

port

Соответсвия полей модели и таблицы в БД

```
[14] df = pd.read_csv('star_classification.csv')
```

```
[16] for it in df.columns:
       print(it)
```

```
obj_ID
alpha
delta
u
g
r
i
z
run_ID
rerun_ID
cam_col
field_ID
spec_obj_ID
class
redshift
plate
MJD
fiber_ID
```

<=>

```
adult=# \d astra3
                    Table "public.astra3
   Column    |        Type        | Collation
-------------+--------------------+-----------
 alpha       | double precision   |
 delta       | double precision   |
 u           | double precision   |
 g           | double precision   |
 r           | double precision   |
 i           | double precision   |
 z           | double precision   |
 run_id      | bigint             |
 cam_col     | bigint             |
 field_id    | bigint             |
 spec_obj_id | double precision   |
 redshift    | double precision   |
 plate       | bigint             |
 mjd         | bigint             |
 fiber_id    | bigint             |

adult=#
```

Соответсвия полей модели и таблицы в БД

```
df.head()
```

	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID
0	1.237661e+18	135.689107	32.494632	23.87882	22.27530	20.39501	19.16573	18.79371	3606	301	2	79	6.543777e+18
1	1.237665e+18	144.826101	31.274185	24.77759	22.83188	22.58444	21.16812	21.61427	4518	301	5	119	1.176014e+19
2	1.237661e+18	142.188790	35.582444	25.26307	22.66389	20.60976	19.34857	18.94827	3606	301	2	120	5.152200e+18
3	1.237663e+18	338.741038	-0.402828	22.13682	23.77656	21.61162	20.50454	19.25010	4192	301	3	214	1.030107e+19
4	1.237680e+18	345.282593	21.183866	19.43718	17.58028	16.49747	15.97711	15.54461	8102	301	3	137	6.891865e+18

```
adult=#  select * from astra3 limit 3;
      alpha       |      delta       |    u     |    g     |    r     |    i     |    z     | run_id | cam_col | field_id |      spec_obj_id
------------------+------------------+----------+----------+----------+----------+----------+--------+---------+----------+-----------------------
 16.9568897845004 | 3.64613008870454 | 23.33542 | 21.95143 | 20.48149 |   19.603 | 19.13094 |   7712 |       6 |      442 |  4.855016555329904e+18 |
 240.063240247767 | 6.13413059813973 | 17.86033 | 16.79228 | 16.43001 | 16.30923 | 16.25873 |   3894 |       1 |      243 | 2.4489280322708705e+18 |
  30.887222067625 | 1.18870964120799 | 18.18911 | 16.89469 | 16.42161 | 16.24627 | 16.18549 |   7717 |       1 |      536 |  8.255357438959835e+18 |
(3 rows)
```

Получение результата:

- В виде таблицы
- Как набор записей из таблицы
- Как набор записей из запроса

Результ таблица

модель

Исходая таблица

```
adult=# SELECT * from  ml_predict_table('astra3.cbm','astra3');
   ml_predict_table
-----------------------
 public.astra3_predict
(1 row)
```

Создается новая таблица

postgres@notebook-sasha: /usr/local/pgsql

```
adult=# select * from astra3_predict;
```

row	alpha	delta	u	g	r	i	z	run_id	cam_col	field_id	spec_obj_id	redshift	plate	mjd	fiber_id	predict	class
1	16.9568897845004	3.64613008870454	23.33542	21.95143	20.48149	19.603	19.13094	7712	6	442	4.855016555329904e+18	0.5062369	4312	55511	495	0.98686	GALAXY
2	240.063240247767	6.13413059813973	17.86033	16.79228	16.43001	16.30923	16.25873	3894	1	243	2.448928032270870 5e+18	0.0003448142	2175	54612	348	).990419	STAR
3	30.887222067625	1.18870964120799	18.18911	16.89469	16.42161	16.24627	16.18549	7717	1	536	8.255357438959835e+18	4.085216e-06	7332	56683	943	).997588	STAR
4	247.594400505002	10.8877797153666	24.99961	21.71203	21.47148	21.30532	21.29109	5323	1	134	4.577998722756271e+18	-0.0002914838	4066	55444	326	).997667	STAR
5	18.8964507920807	-5.26133022886992	23.76648	21.79737	20.69543	20.23403	19.97464	7881	3	148	8.91047176642785e+18	-0.0001361561	7914	57331	363	).996044	STAR
6	182.713733094955	51.3758050594777	22.44608	21.68444	20.24292	19.41423	19.08227	2830	1	411	7.516725588574623e+18	0.5026683	6676	56389	792	).984373	GALAXY
7	150.089423193165	39.4670880748061	18.96441	17.82906	17.31429	16.99891	16.85583	3560	4	278	1.526795411104236e+18	0.06366445	1356	53033	274	).996164	GALAXY
8	189.510984338851	58.7411197772507	21.37376	20.80187	20.84925	21.13449	20.34689	2243	1	353	7.696817897528907e+18	0.7936153	6836	56443	604	).957787	QSO
9	37.7138728560977	-0.525138228146508	20.77988	19.54618	19.16687	18.89438	18.64286	2700	2	117	1.755328312302921 7e+18	0.1060118	1559	53271	183	).993892	GALAXY
10	201.074980072746	28.7699058867715	25.05349	22.23362	20.8122	19.69488	19.28336	4649	3	120	7.306035245308205e+18	0.567082	6489	56329	257	).993856	GALAXY
11	151.83091832672	19.8108624669417	24.04443	22.48608	20.59701	19.50985	19.00457	5183	5	142	6.622787444780849e+18	0.5475619	5882	56029	888	).998885	GALAXY
12	164.364389178099	64.7877852553783	22.98745	23.07199	21.15229	19.97391	19.17876	1302	6	319	7.999579699214046e+18	0.7592601	7105	56740	221	0.99362	GALAXY
13	242.830624949049	52.6659057609914	23.03347	23.60839	21.79315	20.79881	19.74832	1412	4	375	9.081519495399626e+18	0.680221	8066	57544	39	).986432	GALAXY
14	22.9648494083052	0.876034964444952	25.62939	21.3913	19.98634	18.81294	18.13539	4263	6	269	1.213827896058079 2e+18	-1.408066e-05	1078	52643	392	0.99907	STAR
15	176.268553005359	33.432292927676	21.83964	19.87187	18.8392	18.3688	18.09703	4576	1	399	1.155301576648925 6e+19	0.1312984	10261	58462	570	).967815	GALAXY
16	227.751962030348	41.8324497606407	25.98832	23.93829	21.41312	20.34021	19.5577	3664	5	107	9.592801209263348e+18	0.5425886	8520	58191	487	).983403	GALAXY

Результат набор данных

Исходная таблица

Модель

Список категорийных полей

```
adult=#   SELECT * from ml_cat_predict ('titanic.cbm',
'titanic','{name,passenger_id,pclass,sex,sibsp,parch,ticke
t,cabin,embarked }');
 row_num |        predict        | class
---------+-----------------------+-------
       0 |   -1.7937342449233795 | 0
       1 |   -0.7958399022225136 | 0
       2 |    -2.392873216013247 | 0
       3 |    -1.942976624899004 | 0
       4 |  -0.41747860726736713 | 0
       5 |   -2.0608914711097546 | 0
       6 |    0.5914467057444344 | 1
       7 |   -1.0786526230973736 | 0
       8 |    0.6757411102494171 | 1
       9 |    -3.250956928980716 | 0
      10 |    -2.274725588104562 | 0
      11 |   -1.3228896775643357 | 0
      12 |      2.70931909246417 | 1
      13 |   -2.4233542239140187 | 0
```

результат

SELECT * FROM {table}_predict;

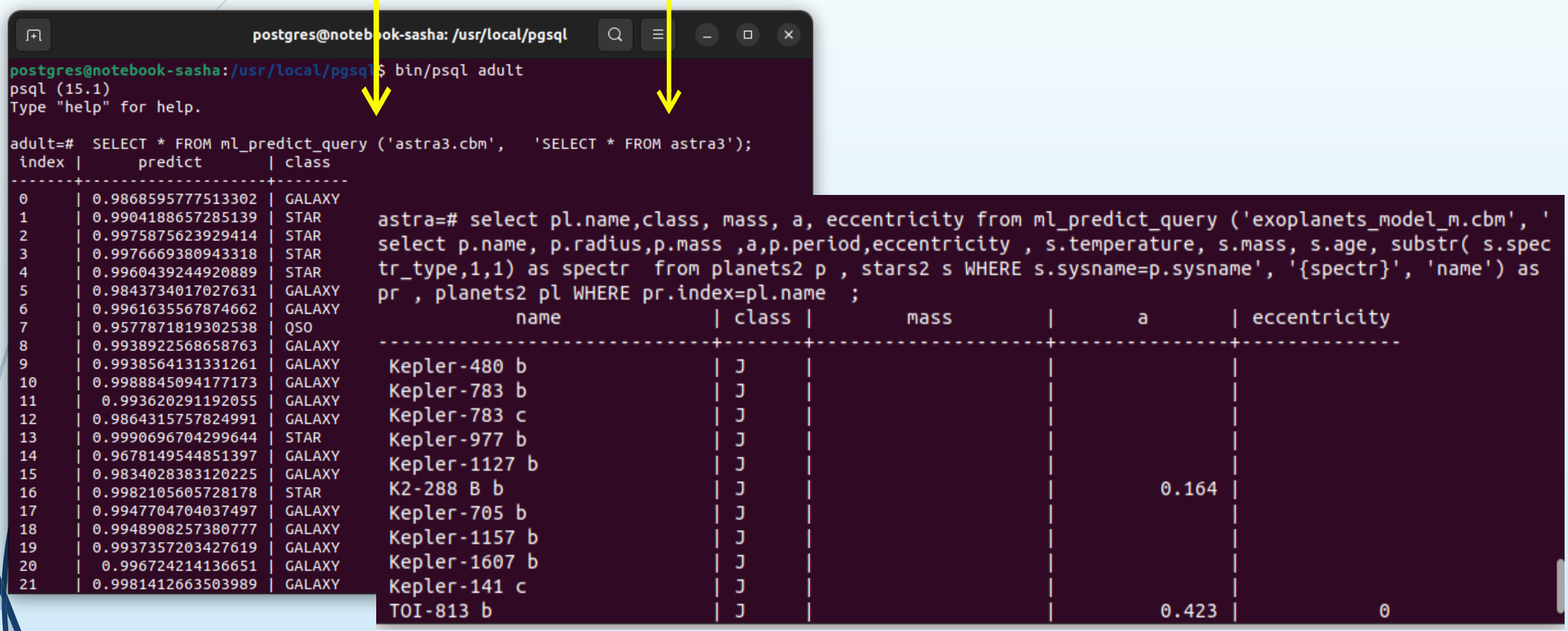
```
adult=# SELECT * from astra3_predict;
 row |       alpha       |       delta       |    u    |    g    |    r    |    i    |    z    | run_id | cam_col | field_id |     spec_obj_id     |   redshift   | plate |  mjd  | fiber_id | predict | class
-----+-------------------+-------------------+---------+---------+---------+---------+---------+--------+---------+----------+---------------------+--------------+-------+-------+----------+---------+--------
   1 |  16.9568897845004 |  3.64613008870454 | 23.3354 | 21.9514 | 20.4814 |  19.603 | 19.1309 |   7712 |       6 |      442 | 4.85501655532990e+18 |    0.5062369 |  4312 | 55511 |      495 | 0.98686 | GALAXY
   2 |  240.063240247767 |  6.13413059813973 | 17.8603 | 16.7922 | 16.4300 | 16.3092 | 16.2587 |   3894 |       1 |      243 | 2.44892803227087e+18 | 0.0003448142 |  2175 | 54612 |      348 | 0.990419 | STAR
   3 |   30.887222067625 |  1.18870964120799 | 18.1891 | 16.8946 | 16.4216 | 16.2462 | 16.1854 |   7717 |       1 |      536 | 8.25535743895983e+18 | 4.085216e-06 |  7332 | 56683 |      943 | 0.997588 | STAR
   4 |  247.594400505002 |  10.8877797153666 | 24.9996 | 21.7120 | 21.4714 | 21.3053 | 21.2910 |   5323 |       1 |      134 | 4.57799872275627e+18 | -0.0002914838 |  4066 | 55444 |      326 | 0.997667 | STAR
   5 |  18.8964507920807 | -5.26133022886992 | 23.7664 | 21.7973 | 20.6954 | 20.2340 | 19.9746 |   7881 |       3 |      148 |  8.9104717664278e+18 | -0.0001361561 |  7914 | 57331 |      363 | 0.996044 | STAR
   6 |  182.713733094955 |  51.3758050594777 | 22.4460 | 21.6844 | 20.2429 | 19.4142 | 19.0822 |   2830 |       1 |      411 | 7.51672558574623e+18 |    0.5026683 |  6676 | 56389 |      792 | 0.984373 | GALAXY
   7 |  150.089423193165 |  39.4670880748061 | 18.9644 | 17.8290 | 17.3142 | 16.9989 | 16.8558 |   3560 |       4 |      278 | 1.52679564111042e+18 |  0.06366445 |  1356 | 53033 |      274 | 0.996164 | GALAXY
   8 |  189.510984338851 |  58.7411197772507 | 21.3737 | 20.8018 | 20.8492 | 21.1344 | 20.3468 |   2243 |       1 |      353 | 7.69681789752890e+18 |    0.7936153 |  6836 | 56443 |      604 | 0.957787 | QSO
   9 |  37.7138728560977 | -0.525138228146508 | 20.7798 | 19.5461 | 19.1668 | 18.8943 | 18.6428 |   2700 |       2 |      117 | 1.75532831230292e+18 |    0.1060118 |  1559 | 53271 |      183 | 0.993892 | GALAXY
  10 |  201.074980072746 |  28.7699058867715 | 25.0534 | 22.2336 |  20.812 | 19.6948 | 19.2833 |   4649 |       3 |      120 | 7.30603524530820e+18 |     0.567082 |  6489 | 56329 |      257 | 0.993856 | GALAXY
  11 |  151.830918326 72 |  19.8108624669417 | 24.0444 | 22.4860 | 20.5970 | 19.5098 | 19.0045 |   5183 |       5 |      142 | 6.62278744478084 9e+18 |    0.5475619 |  5882 | 56029 |      888 | 0.998885 | GALAXY
```

SELECT * FROM ml_predict(...);

```
adult=# SELECT * from  ml_predict('astra3.cbm'
'astra3');
 id  |      predict       | class
-----+--------------------+--------
   0 | 0.9868595777513302 | GALAXY
   1 | 0.9904188657285139 | STAR
   2 | 0.9975875623929414 | STAR
   3 | 0.9976669380943318 | STAR
   4 | 0.9960439244920889 | STAR
   5 | 0.9843734017027631 | GALAXY
   6 | 0.9961635567874662 | GALAXY
   7 | 0.9577871819302538 | QSO
   8 | 0.9938922568658763 | GALAXY
   9 | 0.9938564131331261 | GALAXY
```

Результат, как набор записей из запроса

модель **запрос**



```
postgres@notebook-sasha: /usr/local/pgsql
postgres@notebook-sasha:/usr/local/pgsql$ bin/psql adult
psql (15.1)
Type "help" for help.

adult=#  SELECT * FROM ml_predict_query ('astra3.cbm',   'SELECT * FROM astra3');
 index |      predict       | class
-------+--------------------+--------
 0     | 0.9868595777513302 | GALAXY
 1     | 0.9904188657285139 | STAR
 2     | 0.9975875623929414 | STAR
 3     | 0.9976669380943318 | STAR
 4     | 0.9960439244920889 | STAR
 5     | 0.9843734017027631 | GALAXY
 6     | 0.9961635567874662 | GALAXY
 7     | 0.9577871819302538 | QSO
 8     | 0.9938922568658763 | GALAXY
 9     | 0.9938564131331261 | GALAXY
 10    | 0.9988845094177173 | GALAXY
 11    |  0.993620291192055 | GALAXY
 12    | 0.9864315757824991 | GALAXY
 13    | 0.9990696704299644 | STAR
 14    | 0.9678149544851397 | GALAXY
 15    | 0.9834028383120225 | GALAXY
 16    | 0.9982105605728178 | STAR
 17    | 0.9947704704037497 | GALAXY
 18    | 0.9948908257380777 | GALAXY
 19    | 0.9937357203427619 | GALAXY
 20    |  0.996724214136651 | GALAXY
 21    | 0.9981412663503989 | GALAXY
```

```
astra=# select pl.name,class, mass, a, eccentricity from ml_predict_query ('exoplanets_model_m.cbm', '
select p.name, p.radius,p.mass ,a,p.period,eccentricity , s.temperature, s.mass, s.age, substr( s.spec
tr_type,1,1) as spectr  from planets2 p , stars2 s WHERE s.sysname=p.sysname', '{spectr}', 'name') as
pr , planets2 pl WHERE pr.index=pl.name  ;
            name             | class |        mass        |       a       | eccentricity
-----------------------------+-------+--------------------+---------------+--------------
 Kepler-480 b                | J     |                    |               |
 Kepler-783 b                | J     |                    |               |
 Kepler-783 c                | J     |                    |               |
 Kepler-977 b                | J     |                    |               |
 Kepler-1127 b               | J     |                    |               |
 K2-288 B b                  | J     |                    |         0.164 |
 Kepler-705 b                | J     |                    |               |
 Kepler-1157 b               | J     |                    |               |
 Kepler-1607 b               | J     |                    |               |
 Kepler-141 c                | J     |                    |               |
 TOI-813 b                   | J     |                    |         0.423 |             0
```

SELECT … FROM ml_predict_query(…)

postgres@notebook-sasha: /usr/local/pgsql

```
astra=# SELECT class, predict, name, p_radius,p_mass,metall,p_temperature,a  FROM  ml_predict('exoplanets_model_t.cbm','exoplanets_train_mod
el_by_t' , '{spectr,type_t}', 'name')  p LEFT JOIN  exoplanets_train_model_by_t e ON index=e.name WHERE  class != type_t ;
   class   |       predict       |           name           |  p_radius   |       p_mass       | metall | p_temperature |     a
-----------+---------------------+--------------------------+-------------+--------------------+--------+---------------+-----------
 hot       |  0.5973195076100603 | HD 28254 b               |             |               1.16 |   0.36 |           196 |      2.15
 warm      |  0.6641208386682923 | HD 106252 b              |             |               7.56 | -0.078 |           158 |       2.7
 ice       |    0.95190596023499 | WASP-57 b                |       0.916 |              0.672 |        |          1251 |    0.0386
 hot       |  0.5489876769123071 | HD 147018 c              |             |               6.56 |    0.1 |           171 |     1.922
 warm      |  0.6157999667711632 | HD 202206 c              |             |               2.44 |   0.37 |           159 |      2.55
 hot       |  0.8091935501883233 | HD 5319 b                |             |               1.76 |   0.15 |           303 |    1.6697
 ice       |   0.587509223397627 | DMPP-1 c                 |             |             0.0302 |        |          1239 |    0.0733
 hot       |  0.6296562547910215 | HD 8535 b                |             |               0.68 |   0.02 |           188 |      2.45
 hot       |  0.8156890220280469 | HD 215456 c              |             |              0.246 |        |           178 |     3.394
 hot       |  0.7387209553108713 | Kepler-101 c             |       0.112 |               0.01 |        |          1413 |    0.0684
 warm      | 0.39836805938770775 | HD 86226 b               |             |               0.45 |  0.018 |           176 |      2.73
 hot       |  0.6973761822573443 | Upsilon Andromedae A c   |             |              10.78 |   0.09 |           375 |   0.83158
 hot       |  0.7205908292052307 | Upsilon Andromedae A d   |             |               8.86 |   0.09 |           218 |  2.533539
 hot       |  0.7935630768898031 | Upsilon Andromedae A e   |             |              1.059 |   0.09 |           150 |    5.2456
 very cool |  0.3112231473067454 | Gliese 581 d             |             |              0.019 | -0.135 |           181 |      0.22
 hot       |  0.5547287202957663 | HD 108874 c              |             |              1.018 |   0.14 |           160 |      2.68
 hot       |  0.5930208363415195 | HD 163607 c              |             |               2.29 |   0.21 |           200 |      2.42
 hot       |  0.4981201982814892 | HD 47186 c               |             |            0.35061 |   0.23 |           161 |     2.395
 hot       |  0.8513416828120087 | 2MASS J2002-0521         |        1.36 |                 13 |        |          1301 |
 hot       |  0.9052731374290038 | KOI-0253 c               |     0.10753 |                    |        |           372 |      0.12
 very hot  |  0.8877954989028882 | KOI-4259 b               |     0.10662 |                    |        |          1740 | 0.0171853
```

Информация о модели

```
adult=# SELECT ml_info ('astra3.cbm');
                                     ml_info
-----------------------------------------------------------------------------------
 dimension:3 numeric features:12 categorial features:0 modelType "MultiClass"+
 fieldName:alpha,delta,u,g,r,i,z,cam_col,redshift,plate,MJD,fiber_ID
(1 row)

adult=#
```

- Размерность результата
- Кол-во параметров
- Тип модели
- Имена полей

Информация о моделях

```
adult=# SELECT ml_info ('astra3.cbm');
                                   ml_info
-----------------------------------------------------------------------------
 dimension:3 numeric features:12 categorial features:0 modelType "MultiClass"+
 fieldName:alpha,delta,u,g,r,i,z,cam_col,redshift,plate,MJD,fiber_ID
(1 row)

adult=# SELECT ml_info ('titanic.cbm');
                                            ml_info
-----------------------------------------------------------------------------------------
 dimension:1 numeric features:2 categorial features:9 modelType "Accuracy"          +
 fieldName:PassengerId,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
(1 row)

adult=# SELECT ml_info('boston.cbm');
                                  ml_info
-----------------------------------------------------------------------------
 dimension:1 numeric features:13 categorial features:0 modelType "RMSE"  +
 fieldName:crim,zn,indus,chas,nox,rm,age,dis,rad,tax,ptratio,black,lstat
(1 row)
```

Пример Binary classification

postgres@notebook-sasha: /usr/local/pgsql

```
adult=# select * from titanic_predict;
 row | id  | passenger_id | pclass |                 name                 |  sex   | age  | sibsp | parch |      ticket      |   fare   |    cabin    | embarked | res |  predict  | clas
s
-----+-----+--------------+--------+--------------------------------------+--------+------+-------+-------+------------------+----------+-------------+----------+-----+-----------+-----
--
   1 |   0 |          892 |      3 | Kelly, Mr. James                     | male   | 34.5 |     0 |     0 | 330911           |   7.8292 | -999        | Q        | f   |  0.142616 | 0
   2 |   1 |          893 |      3 | Wilkes, Mrs. James (Ellen Needs)     | female |   47 |     1 |     0 | 363272           |        7 | -999        | S        | f   |  0.310916 | 0
   3 |   2 |          894 |      2 | Myles, Mr. Thomas Francis            | male   |   62 |     0 |     0 | 240276           |   9.6875 | -999        | Q        | f   |  0.083718 | 0
   4 |   3 |          895 |      3 | Wirz, Mr. Albert                     | male   |   27 |     0 |     0 | 315154           |   8.6625 | -999        | S        | f   |  0.125321 | 0
   5 |   4 |          896 |      3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female |   22 |     1 |     1 | 3101298          |  12.2875 | -999        | S        | f   |   0.39712 | 0
   6 |   5 |          897 |      3 | Svensson, Mr. Johan Cervin           | male   |   14 |     0 |     0 | 7538             |    9.225 | -999        | S        | f   |  0.112956 | 0
   7 |   6 |          898 |      3 | Connolly, Miss. Kate                 | female |   30 |     0 |     0 | 330972           |   7.6292 | -999        | Q        | t   |  0.643697 | 1
   8 |   7 |          899 |      2 | Caldwell, Mr. Albert Francis         | male   |   26 |     1 |     1 | 248738           |       29 | -999        | S        | f   |  0.253761 | 0
   9 |   8 |          900 |      3 | Abrahim, Mrs. Joseph (Sophie Halaut Easu) | female |   18 |     0 |     0 | 2657             |   7.2292 | -999        | C        | t   |  0.662787 | 1
  10 |   9 |          901 |      3 | Davies, Mr. John Samuel              | male   |   21 |     2 |     0 | A/4 48871        |    24.15 | -999        | S        | f   |  0.037293 | 0
  11 |  10 |          902 |      3 | Ilieff, Mr. Ylio                     | male   | -999 |     0 |     0 | 349220           |   7.8958 | -999        | S        | f   |  0.093238 | 0
  12 |  11 |          903 |      1 | Jones, Mr. Charles Cresson           | male   |   46 |     0 |     0 | 694              |       26 | -999        | S        | f   |  0.210338 | 0
```

```
adult=#   SELECT * from ml_cat_predict ('titanic.cbm',    'titanic','{name,passenger
_id,pclass,sex,sibsp,parch,ticket,cabin,embarked }');
 row_num |       predict        | class
---------+----------------------+-------
       0 |  -1.7937342449233795 | 0
       1 |  -0.7958399022225136 | 0
       2 |   -2.392873216013247 | 0
       3 |   -1.942976624899004 | 0
       4 | -0.41747860726736713 | 0
       5 |  -2.0608914711097546 | 0
       6 |   0.5914467057444344 | 1
       7 |  -1.0786526230973736 | 0
```

```
adult=# SELECT * FROM ml_cat_predict ('adult.cbm',    'adult2','{workclass,
education,marital_status, occupation,relationship,race,sex,native_country}
');
 row_num |       predict        | class
---------+----------------------+-------
       0 |  -5.926338548423682 | <=50K
       1 |  -1.225876230403332 | <=50K
       2 | -0.7485117670534811 | <=50K
       3 |  3.6351647093731705 | >50K
       4 |  -4.644606242153101 | <=50K
       5 |  -5.342578732065899 | <=50K
       6 | -3.9224526779262296 | <=50K
```

Пример Regression

akalend@notebook-sasha: ~/stars

row	index	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	predict
1	0	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	24.99982
2	1	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	20.664359
3	2	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	33.677379
4	3	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	34.289002
5	4	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	34.615708
6	5	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	27.968317
7	6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	21.682186
8	7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	22.853984
9	8	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	17.011092
10	9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	18.24062
11	10	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	17.543837
12	11	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	19.974764
13	12	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	20.70866
14	13	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	20.202922
15	14	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	18.175456
16	15	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9	19.698336
17	16	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1	22.304514
18	17	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5	17.160698
19	18	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2	18.706903
20	19	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2	18.769804
21	20	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6	13.985032

--Далее--

```
adult=#   SELECT * from ml_cat_predict ('boston.cbm',  'boston2');
 row_num |      predict       | class
---------+--------------------+-------
       0 |  24.99982028068538 |
       1 | 20.664358727562394 |
       2 |  33.67737911788664 |
       3 |  34.28900239364565 |
       4 |  34.61570849423551 |
       5 | 27.968317495475695 |
       6 |  21.68218578618033 |
```

Пример Multi classification

```
adult=# SELECT * from astra3_predict;
 row |      alpha       |      delta       |    u    |    g    |    r    |    i    |    z    | run_id | cam_col | field_id |      spec_obj_id      |   redshift    | plate | mjd | fiber_id | predict | class
-----+------------------+------------------+---------+---------+---------+---------+---------+--------+---------+----------+-----------------------+---------------+-------+-----+----------+---------+--------
   1 | 16.9568897845004 |  3.64613008870454 | 23.33542 | 21.95143 | 20.48149 |  19.603 | 19.13094 |   7712 |       6 |      442 | 4.855016555329904e+18 |     0.5062369 |  4312 | 55511 |      495 | 0.98686 | GALAXY
   2 | 240.063240247767 |  6.13413059813973 | 17.86033 | 16.79228 | 16.43001 | 16.30923 | 16.25873 |   3894 |       1 |      243 | 2.4489280322708705e+18 |  0.0003448142 |  2175 | 54612 |      348 | .990419 | STAR
   3 |  30.887222067625 |  1.18870964120799 | 18.18911 | 16.89469 | 16.42161 | 16.24627 | 16.18549 |   7717 |       1 |      536 | 8.255357438959835e+18 | 4.085216e-06 |  7332 | 56683 |      943 | .997588 | STAR
   4 | 247.594400505002 |  10.8877797153666 | 24.99961 | 21.71203 | 21.47148 | 21.30532 | 21.29109 |   5323 |       1 |      134 | 4.577998722756271e+18 | -0.0002914838 |  4066 | 55444 |      326 | .997667 | STAR
   5 | 18.8964507920807 | -5.26133022886992 | 23.76648 | 21.79737 | 20.69543 | 20.23403 | 19.97464 |   7881 |       3 |      148 | 8.91047176642785e+18 | -0.0001361561 |  7914 | 57331 |      363 | .996044 | STAR
   6 | 182.713733094955 |  51.3758050594777 | 22.44608 | 21.68444 | 20.24292 | 19.41423 | 19.08227 |   2830 |       1 |      411 | 7.516725588574623e+18 |     0.5026683 |  6676 | 56389 |      792 | .984373 | GALAXY
   7 | 150.089423193165 |  39.4670880748061 | 18.96441 | 17.82906 | 17.31429 | 16.99891 | 16.85583 |   3560 |       4 |      278 | 1.5267956411104236e+18 |    0.06366445 |  1356 | 53033 |      274 | .996164 | GALAXY
   8 | 189.510984338851 |  58.7411197772507 | 21.37376 | 20.80187 | 20.84925 | 21.13449 | 20.34689 |   2243 |       1 |      353 | 7.696817897528907e+18 |     0.7936153 |  6836 | 56443 |      604 | .957787 | QSO
   9 | 37.7138728560977 | -0.525138228146508 | 20.77988 | 19.54618 | 19.16687 | 18.89438 | 18.64286 |   2700 |       2 |      117 | 1.7553283123029217e+18 |     0.1060118 |  1559 | 53271 |      183 | .993892 | GALAXY
  10 | 201.074980072746 |  28.7699058867715 | 25.05349 | 22.23362 |  20.8122 | 19.69488 | 19.28336 |   4649 |       3 |      120 | 7.306035245308205e+18 |      0.567082 |  6489 | 56329 |      257 | .993856 | GALAXY
  11 | 151.8309183672 |  19.8108624669417 | 24.04443 | 22.48608 | 20.59701 | 19.50985 | 19.00457 |   5183 |       5 |      142 | 6.622787444780849e+18 |     0.5475619 |  5882 | 56029 |      888 | .998885 | GALAXY
```

```
adult=# SELECT * from  ml_predict('astra3.cbm'
'astra3');
 id |       predict       | class
----+---------------------+--------
  0 | 0.9868595777513302  | GALAXY
  1 | 0.9904188657285139  | STAR
  2 | 0.9975875623929414  | STAR
  3 | 0.9976669380943318  | STAR
  4 | 0.9960439244920889  | STAR
  5 | 0.9843734017027631  | GALAXY
  6 | 0.9961635567874662  | GALAXY
  7 | 0.9577871819302538  | QSO
  8 | 0.9938922568658763  | GALAXY
  9 | 0.9938564131331261  | GALAXY
```

```
astra=# SELECT class, predict, index as name  FROM  ml_predict('exoplanets_
model_t.cbm','exoplanets_train_model_by_t' , '{spectr,type_t}', 'name');
   class    |       predict       |            name
------------+---------------------+----------------------------
 hot        |  0.9836827156493978 | Kepler-89 d
 hot        |  0.5973195076100603 | HD 28254 b
 ice        |  0.9732892668460026 | WASP-28 b
 hot        |  0.9989161355601188 | WASP-84 b
 warm       |  0.6641208386682923 | HD 106252 b
 ice        |  0.9940189501044985 | Kepler-8 b
 ice        |    0.95190596023499 | WASP-57 b
 hot        |  0.7579202916405879 | HD 173416 b
 hot        |  0.9981612189077735 | WASP-29 b
 ice        |  0.9245818348486754 | HD 3167 b
 hot        |   0.976449070282651 | K2-240 b
 ice        |  0.9949239070889919 | NGTS-24 b
 hot        |  0.9981352494125506 | TOI-421 b
 very hot   |  0.9609388933704017 | KELT-7 b
 hot        |  0.7460383200300617 | K2-16 c
 hot        |  0.5489876769123071 | HD 147018 c
 warm       |  0.6157999667711632 | HD 202206 c
 hot        |  0.8091935501883233 | HD 5319 b
```

Внутренние данные

```
adult=#
adult=# select name, j #> '{data_processing_options,cla
  name   |            class             |  loss_func
---------+--------------------------------+--------------
 astra   | ["GALAXY", "QSO", "STAR"] | "MultiClass"
 titanic | [0, 1]                         | "Logloss"
 titanic | [0, 1]                         | "Logloss"
 boston  | []                             | "RMSE"
 adult   | ["<=50K", ">50K"]              | "Logloss"
(5 rows)
```

PostgeSQL и ClickHouse

```
postgres@notebook-sasha: /usr/local/pgsql

adult=# SELECT * FROM  ml_predict ('titanic.cbm', 'titanic','{name,passen
ger_id,pclass,sex,sibsp,parch,ticket,cabin,embarked }', 'passenger_id') L
IMIT 10;
 index |       predict        | class
-------+----------------------+-------
 892   |  0.14261550408912857 | 0
 893   |   0.3109161066836635 | 0
 894   |  0.08371776629817723 | 0
 895   |  0.12532120732529184 | 0
 896   |   0.3971202544185592 | 0
 897   |  0.11295647623974053 | 0
 898   |   0.6436970182793165 | 1
 899   |  0.25376107990544644 | 0
 900   |   0.6627874954602727 | 1
 901   | 0.037292516669765206 | 0
(10 rows)

adult=#
```

Множественная классификация не поддерживается

```
akalend@notebook-sasha: ~

notebook-sasha :) SELECT PassengerId, catboostEvaluate('/tmp/titanic.cb
m',
 PassengerId, Pclass, Name, Sex, Age, SibSp,
 Parch, Ticket,
 Fare, Cabin, Embarked) AS prediction
FROM titanic
LIMIT 10

SELECT
    PassengerId,
    catboostEvaluate('/tmp/titanic.cbm', PassengerId, Pclass, Name, Sex
, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked) AS prediction
FROM titanic
LIMIT 10

Query id: cb0c6ac1-363e-4f4d-a172-54c06b6659b8

┌─PassengerId─┬──────────prediction─┐
│         892 │  -1.993659010153981 │
│         893 │   -2.15638521773551 │
│         894 │ -2.0175711183142973 │
│         895 │   -2.15638521773551 │
│         896 │ -2.0847733562958295 │
│         897 │   -2.15638521773551 │
│         898 │  -1.993659010153981 │
│         899 │  -2.108685464456145 │
│         900 │ -1.8606454378897987 │
│         901 │   -2.15638521773551 │
└─────────────┴─────────────────────┘

← Progress: 418.00 rows, 50.62 KB (33.25 thousand rows/s., 4.03 MB/s.)
↖ Progress: 418.00 rows, 50.62 KB (33.25 thousand rows/s., 4.03 MB/s.)

10 rows in set. Elapsed: 0.013 sec.

notebook-sasha :)
```

сравнение
pg_ml и PostgesML

XGBoost

```
                    Table "public.titanic"
   Column    |       Type       | Collation |
-------------+------------------+-----------+
id           | integer          |           |
passenger_id | integer          |           |
pclass       | integer          |           |
name         | text             |           |
sex          | text             |           |
age          | double precision |           |
sibsp        | integer          |           |
parch        | integer          |           |
ticket       | text             |           |
fare         | double precision |           |
cabin        | text             |           |
embarked     | character(1)     |           |
res          | boolean          |           |
```

On hot coding

```
postgresml=# \d titanic
                    Table "public.titanic"
     Column     |       Type       | Collation
----------------+------------------+-----------
 index          | bigint           |
 unnamed:0      | bigint           |
 pclass         | bigint           |
 name           | text             |
 sib_sp         | bigint           |
 parch          | bigint           |
 ticket         | text             |
 fare           | double precision |
 sex_male       | bigint           |
 nulls_1        | bigint           |
 nulls_2        | bigint           |
 cabin_mapped_1 | bigint           |
 cabin_mapped_2 | bigint           |
 cabin_mapped_3 | bigint           |
 cabin_mapped_4 | bigint           |
 cabin_mapped_5 | bigint           |
 cabin_mapped_6 | bigint           |
 cabin_mapped_7 | bigint           |
 cabin_mapped_8 | bigint           |
 embarked_q     | bigint           |
 embarked_s     | bigint           |
 survived       | bigint           |
Indexes:
    "ix_titanic_index" btree (index)
```

Использовались наборы данных

1. Titanic https://www.kaggle.com/datasets/heptapod/titanic
2. Adult
   https://www.kaggle.com/datasets/brijeshbmehta/adult-datasets/data
3. Predicting Pulsar star
   https://www.kaggle.com/datasets/colearninglounge/predicting-pulsar-starintermediate/discussion
4. NASA Exoplanet catalog
   https://exoplanetarchive.ipac.caltech.edu/

Спасибо за внимание

Приглашаются желающие в проект

Вопросы?