СБЕР УНИВЕРСИТЕТ

Дипломный проект на тему:
«Сравнительный анализ работы алгоритмов ML и DL»

Выполнил слушатель группы AR
Куделькин А.Е.

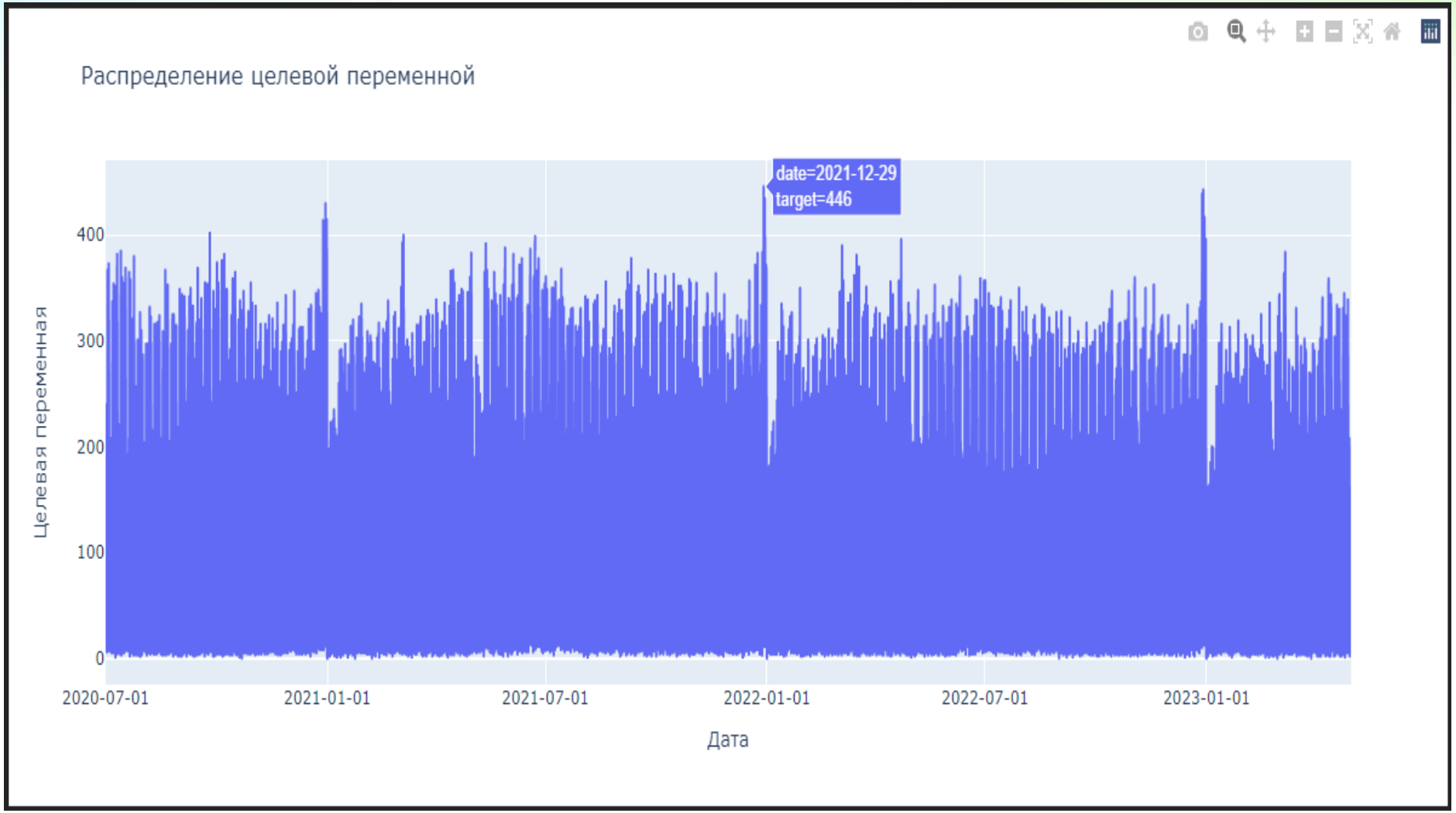
СБЕР УНИВЕРСИТЕТ

Цели проекта:

- Провести эксперимент по прогнозированию бизнес драйвера крупного ритейлера классическими алгоритмами и рекуррентной нейронной сетью(LSTM).
- Подобрать наилучшую модель и гиперпараметры используя метрику качества – коэффициент детерминации.
- Оптимизировать работу модели для наилучшего прогноза январских праздничных дней

СБЕР УНИВЕРСИТЕТ

Количество продаваемого товара в часовой грануле



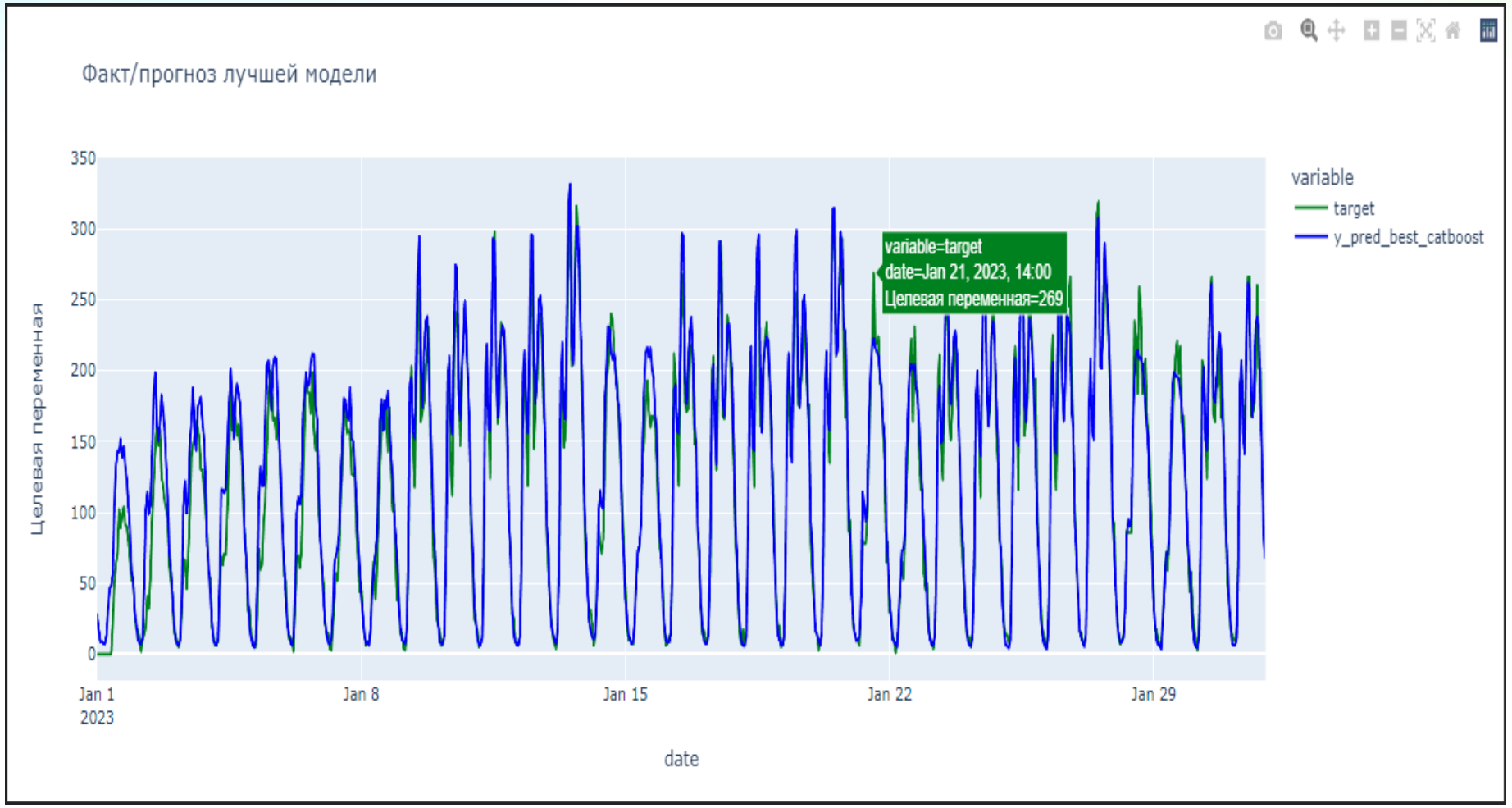
Выбор алгоритма и подбор гиперпараметров по сетке

1. CatBoost – 0.95837
2. GradientBoosting – 0.95185
3. XGBoost – 0.94705
4. RandomForest – 0.92917

```
Модель CatBoost:
Fitting 5 folds for each of 14 candidates, totalling 70 fits
Лучшие параметры: {'forecast__depth': 6, 'forecast__learning_rate': 0.1}
МЕТРИКИ МОДЕЛИ:
MAE: 13.64812
MAPE: 15.07046
MSE: 377.17989
RMSE: 19
R2: 0.95837
--------------------------------------------------------------------------------
CPU times: total: 9.72 s
Wall time: 36.3 s

Модель XGBoost:
Fitting 5 folds for each of 14 candidates, totalling 70 fits
Лучшие параметры: {'forecast__learning_rate': 0.1, 'forecast__max_depth': 7}
МЕТРИКИ МОДЕЛИ:
MAE: 14.45363
MAPE: 14.74073
MSE: 459.19475
RMSE: 21
R2: 0.94705
--------------------------------------------------------------------------------
CPU times: total: 3.75 s
Wall time: 8.59 s

Модель GradientBoostingRegressor:
Fitting 5 folds for each of 21 candidates, totalling 105 fits
Лучшие параметры: {'forecast__max_depth': 5, 'forecast__n_estimators': 300}
МЕТРИКИ МОДЕЛИ:
MAE: 14.44979
MAPE: 14.63471
MSE: 419.35807
RMSE: 20
R2: 0.95185
--------------------------------------------------------------------------------
CPU times: total: 8.16 s
Wall time: 1min 48s

Модель RandomForest:
Fitting 5 folds for each of 21 candidates, totalling 105 fits
Лучшие параметры: {'forecast__max_depth': 7, 'forecast__n_estimators': 100}
МЕТРИКИ МОДЕЛИ:
MAE: 15.94111
MAPE: 16.95827
MSE: 585.00831
RMSE: 24
R2: 0.92917
--------------------------------------------------------------------------------
CPU times: total: 3.19 s
Wall time: 1min 20s
CPU times: total: 24.8 s
Wall time: 3min 54s
```

СБЕР УНИВЕРСИТЕТ

Результат работы на январе 2023г.



Плохой результат на первых днях года и отличные показатели на остальном месяце

СБЕР УНИВЕРСИТЕТ

Небольшая переобученность, но надо учитывать сложность контрольной выборки

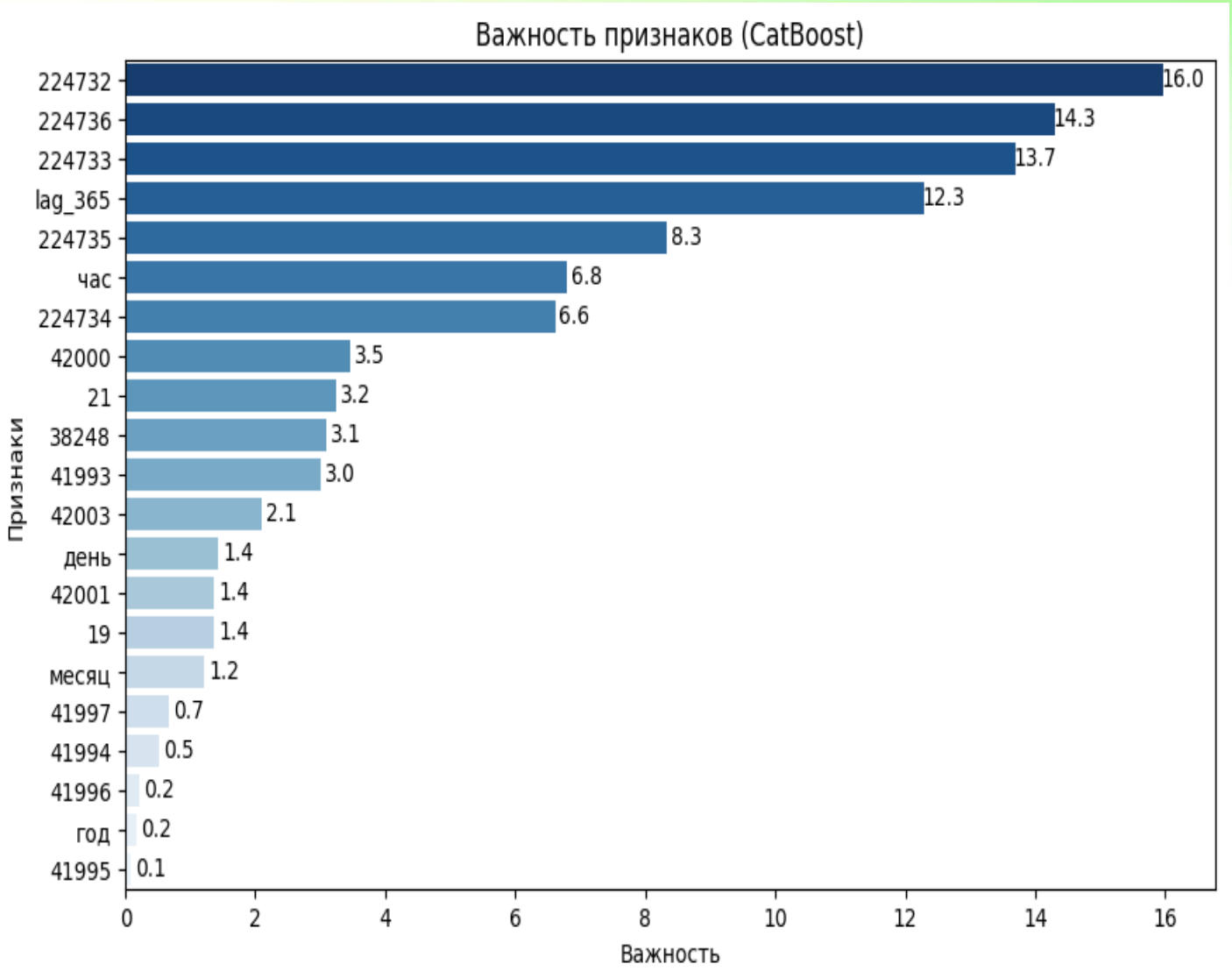
```
Модель CatBoost:
Fitting 5 folds for each of 14 candidates, totalling 70 fits
Лучшие параметры: {'forecast__depth': 6, 'forecast__learning_rate': 0.1}
МЕТРИКИ МОДЕЛИ:
MAE: 13.64812
MAPE: 15.07046
MSE: 377.17989
RMSE: 19
R2: 0.95837
--------------------------------------------------------------------------
```

```
МЕТРИКИ НА КОНТРОЛЬНОЙ ВЫБОРКЕ:
MAE: 14.78451
MAPE: 18.9119
MSE: 437.77204
RMSE: 21
R2: 0.93968
--------------------------------------------------------------------------
```

	R2
1	0.128900
2	0.407100
3	0.703500
4	0.785500
5	0.727700
6	0.844800
7	0.925400
8	0.893800
9	0.933300
10	0.956200
11	0.969200
12	0.951800
13	0.976100
14	0.936700
15	0.881300
16	0.947500
17	0.978600
18	0.970700
19	0.945300
20	0.969500
21	0.960500
22	0.953500
23	0.965000
24	0.985000
25	0.961900
26	0.954400
27	0.989500
28	0.939200
29	0.969700
30	0.962100
31	0.971200

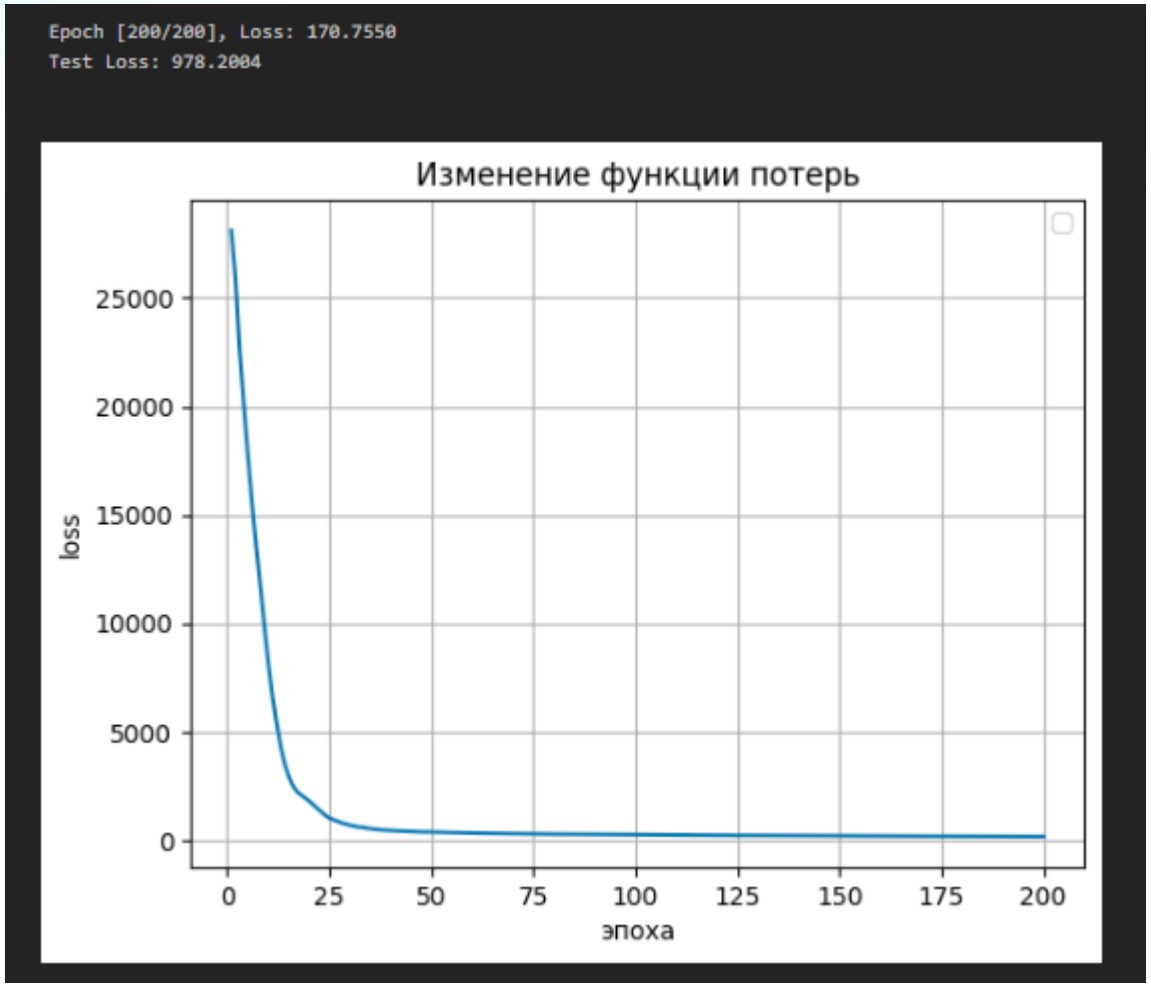
СБЕР УНИВЕРСИТЕТ

Наиболее влиятельными признаками являются Лаги с 35 по 91 день с шагом 7, для лучшего контроля недельного профиля



RNN с архитектурой LSTM

Несмотря на использование регуляризатора dropout, нейронная сеть сильно переобучается

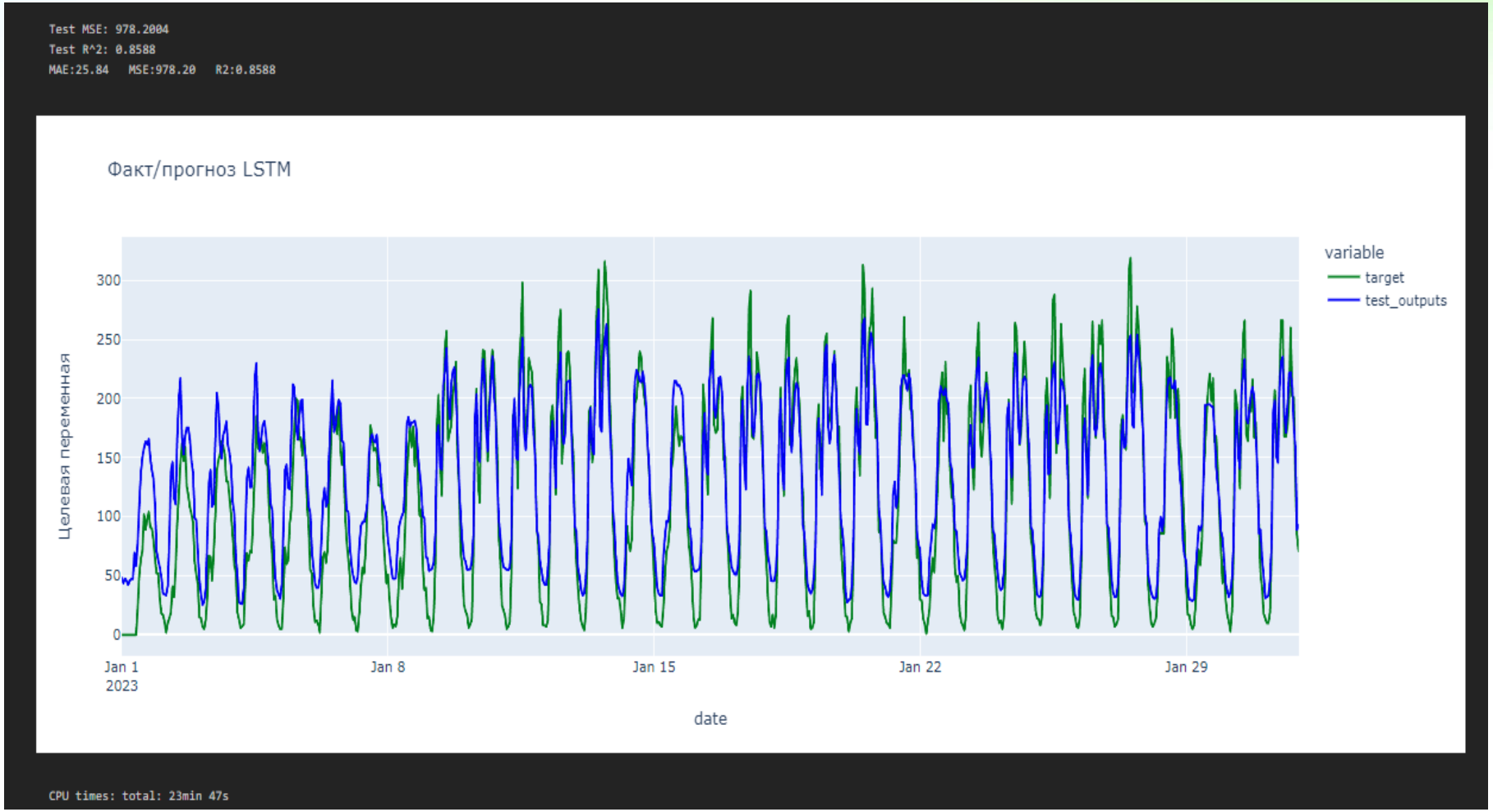


Параметр	Значение
input_size	21
hidden_size	400
num_layers	2
output_size	1
num_epochs	200
learning_rate	0.07
dropout_prob	0.2
criterion	MSELoss
optimizer	Adam

СБЕР УНИВЕРСИТЕТ

Результат работы LSTM

Плохие результаты в начале года + неверные прогнозы на нижних и верхних значениях



СБЕР УНИВЕРСИТЕТ

Сравнительный анализ работы ML и DL

LSTM

	R2
1	-0.813700
2	-0.159200
3	0.395500
4	0.508700
5	0.567200
6	0.750000
7	0.671600
8	0.578300
9	0.868700
10	0.900600
11	0.898300
12	0.914200
13	0.924600
14	0.858200
15	0.756300
16	0.897500
17	0.893200
18	0.915100
19	0.936500
20	0.939200
21	0.912900
22	0.861500
23	0.900600
24	0.938200
25	0.895100
26	0.927800
27	0.928100
28	0.907000
29	0.916000
30	0.928900
31	0.936100

CatBoost

	R2
1	0.128900
2	0.407100
3	0.703500
4	0.785500
5	0.727700
6	0.844800
7	0.925400
8	0.893800
9	0.933300
10	0.956200
11	0.969200
12	0.951800
13	0.976100
14	0.936700
15	0.881300
16	0.947500
17	0.978600
18	0.970700
19	0.945300
20	0.969500
21	0.960500
22	0.953500
23	0.965000
24	0.985000
25	0.961900
26	0.954400
27	0.989500
28	0.939200
29	0.969700
30	0.962100
31	0.971200

При работе с табличными данными, в виде временных рядов, можно выделить следующее:

1.Существенный перевес в сторону классических алгоритмов ML по времени прогнозирования;
2. Точность прогноза ощутимо выше у любого подвида градиентного бустинга при одинаково затраченном времени подбора гиперпараметров;
3. Требования к вычислительным мощностям также намного скромнее у классических моделей.

Проект выполнен на CPU Intel Core i7-10700/32Gb, что недостаточно для быстроты работы нейронных сетей. Возможно с GPU удалось бы подобрать гиперпараметры и получить более лучшие результаты.

СБЕР УНИВЕРСИТЕТ