

;>

Aeroclub Challenge 2023

12 - 27 мая



ML_VODOLAZEZ

Создание сервиса ранжирования

предложений Auto Avia Offer

Track# 2

ML_V0d0lazezz



Илларионов Алексей

ML-engineer



Голубкина Анна

Data-Scientist



Юдаков Евгений

Менеджер проекта



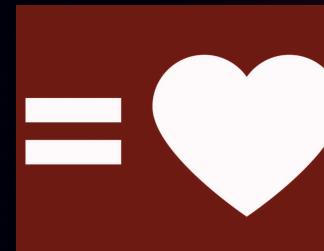
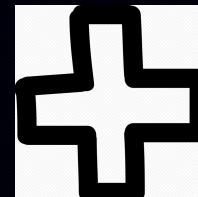
Описание задачи

На входе модель получает xls-файл с возможными вариантами перелета по запрошеному клиентом маршруту.

В результате работы модель сортирует загруженные варианты в порядке уменьшения приоритета предложения, от 1 до N, где N - номер наименее подходящего варианта

$$L = \frac{v_0^2 \sin^2 \alpha}{g}$$

Дальность_полёта

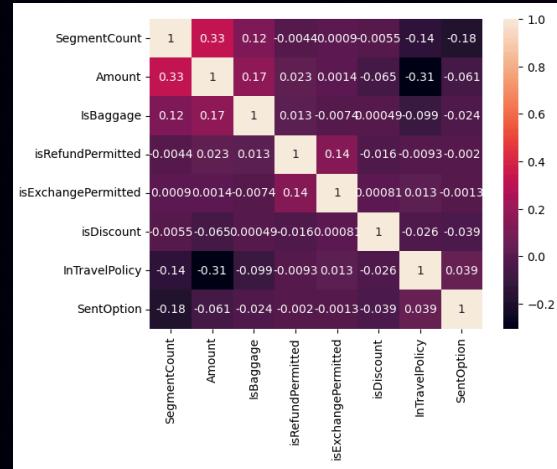


Сбор и обработка данных

- 1) Заполнены пропущенные данные
- 2) Даты приведены к нулевому меридиану для возможности расчета времени в пути
- 3) Введены дополнительные временные признаки
- 3) Список авиакомпаний выведен, как отдельная фича
- 4) Введен новый признак «разница между запрошенным временем и фактическим временем рейса»

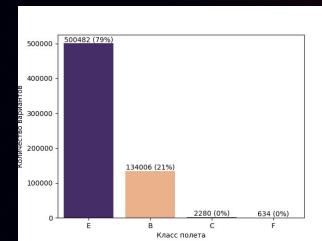
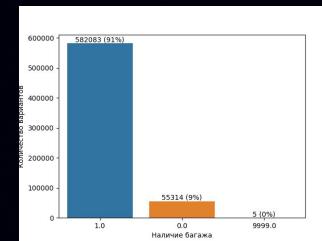
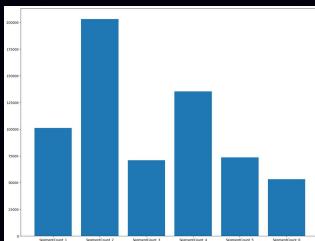
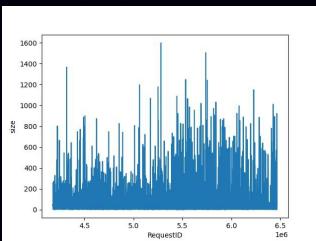
Пример:

Клиенту необходимо прибыть в локацию в 16:00 сегодня. Модель считает приоритетными рейсы с близким временем отправления и точным расчетом прибытия до 16:00 в соответствующую локацию.



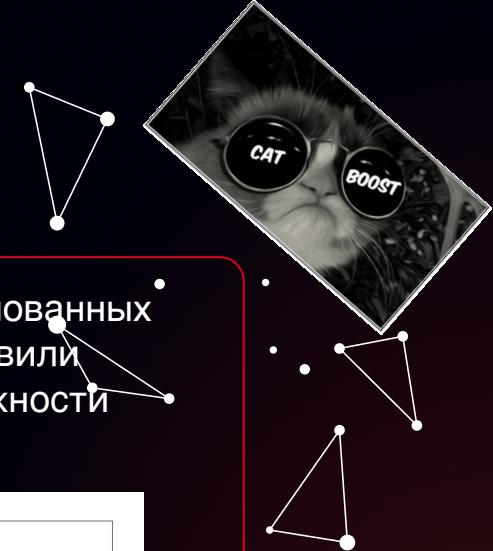
Выбор модели

Был рассмотрен вариант использования рекомендательных систем основанных на применении алгоритма Alternating least squares (ALS), однако остановили выбор на модели градиентного бустинга по причине отсутствия возможности использования клиентского опыта.



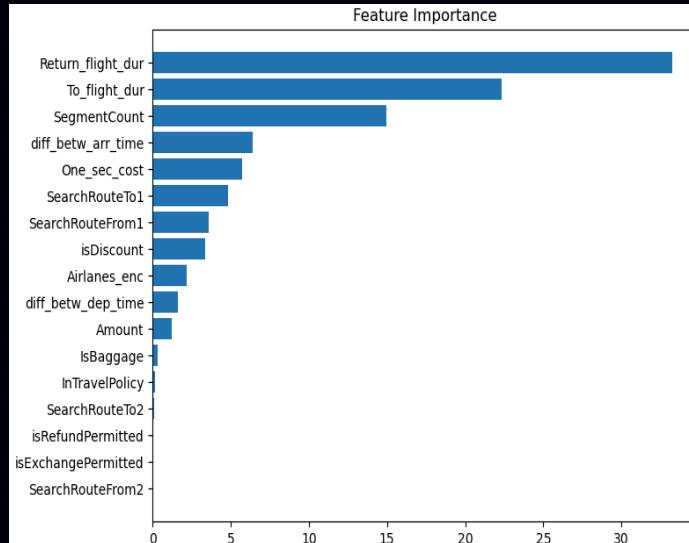
Поэтому выбрана модель CatBoost

CatBoost — открытая программная библиотека, разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга.



Обучение модели

опишите процесс обучения модели и как вы оптимизировали ее параметры
CatBoostClassifier с метрикой оценки f1-score



```
In [*]: ctb_clsf_model = CatBoostClassifier(random_seed=42, cat_features=pipeline.cat_feats, silent=True)

grid = {'learning_rate': [0.03, 0.1],
        'depth': [4, 6, 10],
        'l2_leaf_reg': [1, 3, 5, 7, 9]}

grid_search_result = ctb_clsf_model.grid_search(grid,
                                                X=pipeline.df,
                                                y=pipeline.target,
                                                plot=True, )
```

bestTest = 0.1126779175
bestIteration = 999

0: loss: 0.1126779 best: 0.1126779 (0) total: 7m 27s remaining: 3h 36m 5s

bestTest = 0.1086485935
bestIteration = 999

1: loss: 0.1086486 best: 0.1086486 (1) total: 15m 42s remaining: 3h 39m 48s

bestTest = 0.1129513874
bestIteration = 999

2: loss: 0.1129514 best: 0.1086486 (1) total: 22m 32s remaining: 3h 22m 49s

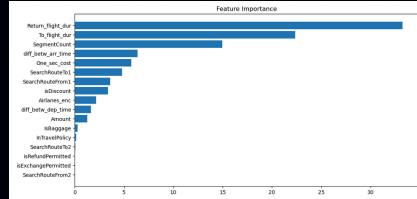


Архитектура



Результаты

Благодаря формированию новых признаков мы можем посчитать все необходимые атрибуты, предложив максимально подходящее предложение.



Модель прекрасно обучилась и мы получили хорошую точность.

```
[113]: print(classification_report(test_pipeline.target, predict))
```

	precision	recall	f1-score	support
0	0.97	1.00	0.98	55329
1	0.33	0.04	0.07	2028
accuracy			0.96	57357
macro avg	0.65	0.52	0.52	57357
weighted avg	0.94	0.96	0.95	57357

Ограничения у данной модели это необходимость передачи входных данных по строгому шаблону. Т.е. при изменении названий полей работа модели будет остановлена с ошибкой.

r.num	5342	5355	5350	5345	5347	5356	5353
r.id	5343	5343	5356	5351	5346	5348	5357
r.requestid	6410888	6410888	6410888	6410888	6410888	6410888	6410888
r.employeeid	1461	1461	1461	1461	1461	1461	1461
r.requestdate	2022-12-20 14:09:30	2022-12-20 14:09:30	2022-12-20 14:09:30	2022-12-20 14:09:30	2022-12-20 14:09:30	2022-12-20 14:09:30	2022-12-20 14:09:30
r.edition	45112	45112	45112	45112	45112	45112	45112
r.volumen							
r.searchroute	AEREVN/EVNAER						
r.requestedpartner	2022-12-30 00:00:00	2022-12-30 00:00:00	2022-12-30 00:00:00	2022-12-30 00:00:00	2022-12-30 00:00:00	2022-12-30 00:00:00	2022-12-30 00:00:00
r.requestreturndate	2023-01-06 00:00:00	2023-01-06 00:00:00	2023-01-06 00:00:00	2023-01-06 00:00:00	2023-01-06 00:00:00	2023-01-06 00:00:00	2023-01-06 00:00:00
r.flagoption	A40829 AEREVN 2022						
r.departuredate	2022-12-30 08:25:00	2022-12-30 08:25:00	2022-12-30 08:25:00	2022-12-30 08:25:00	2022-12-30 08:25:00	2022-12-30 08:25:00	2022-12-30 08:25:00
r.arrivaldate	2022-12-30 10:55:00	2022-12-30 10:55:00	2022-12-30 10:55:00	2022-12-30 10:55:00	2022-12-30 10:55:00	2022-12-30 10:55:00	2022-12-30 10:55:00
r.returndepaturedate	2023-01-06 11:45:00	2023-01-06 11:45:00	2023-01-06 11:45:00	2023-01-06 11:45:00	2023-01-06 11:45:00	2023-01-06 11:45:00	2023-01-06 11:45:00
r.returnarrivedate	2023-01-06 11:50:00	2023-01-06 11:50:00	2023-01-06 11:50:00	2023-01-06 11:50:00	2023-01-06 11:50:00	2023-01-06 11:50:00	2023-01-06 11:50:00
r.segmentcount	2	2	2	2	2	2	2
r.amount	40593	36728	111523	32223	40593	50919	50919
r.class	E	E	E	E	E	E	E
r.isbaggage	1	0	1	0	1	1	1
r.isrefundpermitted	1	0	1	0	1	1	1
r.istravelerpermit	1	1	1	1	1	1	1
r.discount	0	0	0	0	0	0	0
r.intravelpolicy	1	1	1	1	1	1	1
r.position (from 1 to n)							
r.target	37%	34%	29%	23%	13%	11%	11%
row_number_windic	1	2	3	4	5	7	6



Выводы

Модель по итогу показала хороший результат, но его можно улучшить детальнее проработать входные данные, добавив новые признаки.

Поработать с балансировкой данных, так как в классах наблюдается значительный дисбаланс.

Благодаря более детальному анализу данных мы можем принять во внимание больше признаков, а соответственно и улучшить качество работы модели.

idnum	4336	4337	2535	2530	2531
r.id	4337	4338	2536	2531	
r.requestid	5852970	5852970	5852970	5852970	
r.destinationid	125	125	1568	1568	
r.requestdate	01.11.2022 13:09	01.11.2022 13:09	12.10.2022 14:37	12.10.2022 14:37	
r.clientid	24820	24820	54545	54545	
r.valueuru					
r.searchroute	CSYLED	CSYLED	ALAMOW	ALAMOW	
r.requestdeparturedate	11.11.2022 0:00	11.11.2022 0:00	29.10.2022 0:00	29.10.2022 0:00	
r.requestreturndate					
r.flightoption	DP0526 CSYLED 2022.11.11	DP0526 CSYLED 2022.11.11	DV0815 ALAVKO 2022.10.29	DV0815 ALAVKO 2022.10.29	
r.departuredate	11.11.2022 20:30	11.11.2022 20:30	29.10.2022 19:30	29.10.2022 19:30	
r.arrivaldate	11.11.2022 22:40	11.11.2022 22:40	29.10.2022 21:56	29.10.2022 20:50	
r.returndestinationsdate					
r.returninvaliddate					
r.segmentcount	1	1	1	1	
r.amount	4074	5324	19781	21216	
r.class	E	E	E	E	
r.baggage	0	1	1	1	
r.petsgoing	1	1	0	0	
r.petscomingpermitted	1	1	1	1	
r.discount	0	0	0	0	
c.intravelpolicy	1	1	1	1	
r.position (from 1 to n)					
r.target	65%	62%	61%	58%	
row_number_window_0	1	2	1	2	

