

Кейс от Ростелекома

Команда @sergak_blog

Проблематика



Министерство экономического развития
Российской Федерации



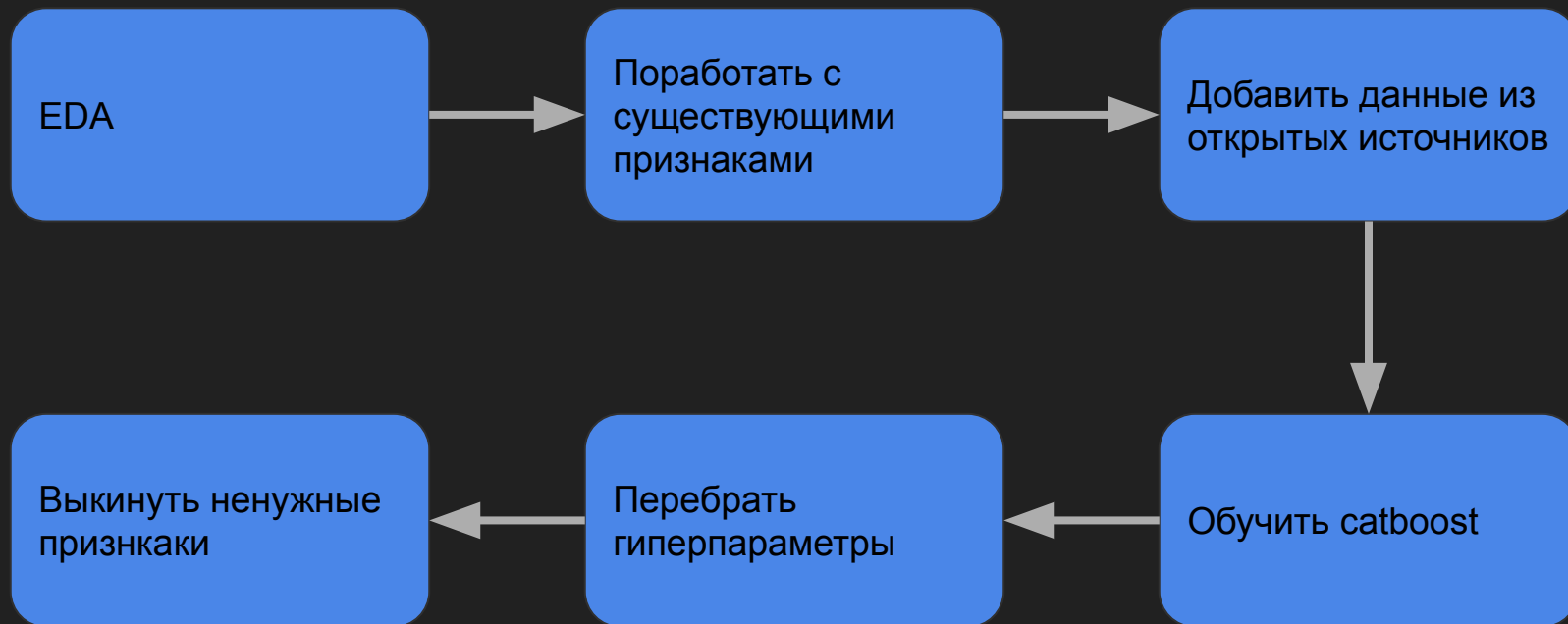
цифровой
прорыв

сезон: III

Проблематика

Сбор и проверка данных очень длительный и трудоемкий процесс (не говоря уже о внедрении всего этого в промышленный пайплайн).

Общий пайплайн



Работаем с существующими признаками

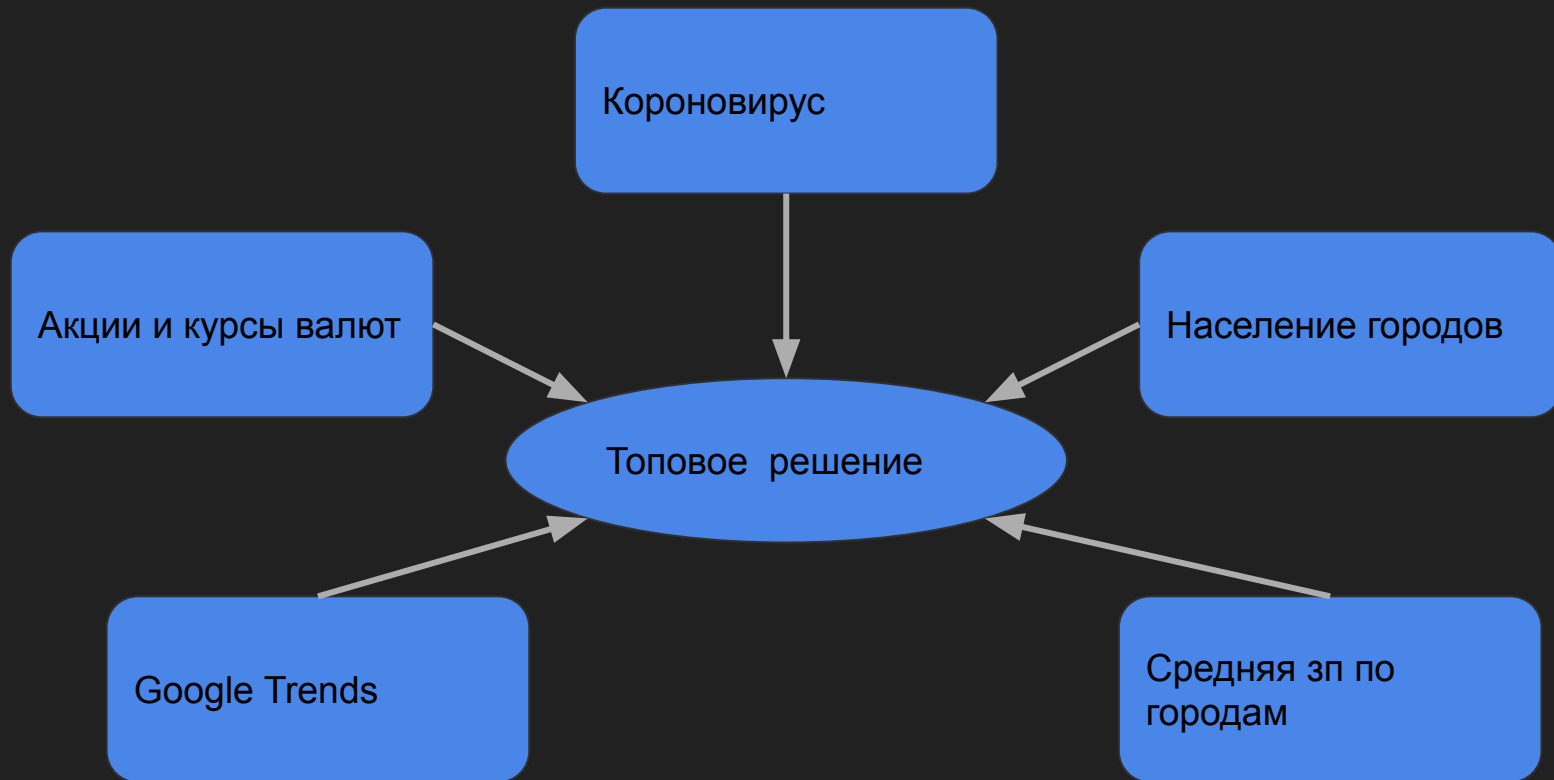
- Сделать агрегации по городам для f1-f30
- Посчитать парные значения для f1-f30 (разность, отношение)
- Нормализовать названия городов/регионов (.lower().strip())

```
data = pd.read_csv('data/train.csv', sep=';')
data['city_name'] = data['city_name'].str.strip().str.lower()
data.head()
```

	label	period	subject_type	subject_name	city_name	hex	hex_lat	hex_lon	f1	f2	...	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30
0	1	2020-05-01	Город	Москва	москва	8611aa7a7ffffff	55.729458	37.516569	0.00101	0.00103	...	0.13027	0.0	0.00000	NaN	NaN	NaN	NaN	NaN	0.01737	0.0
1	1	2020-05-01	Город	Москва	москва	8611aa01ffffff	55.975851	37.237085	0.00000	0.00027	...	0.08756	0.0	0.00000	NaN	NaN	NaN	NaN	NaN	0.01152	0.0
2	1	2020-05-01	Город	Москва	москва	861181b6ffffff	55.622721	37.695121	0.00339	0.00313	...	0.09243	0.0	0.00000	0.11053	0.57895	0.00526	0.02105	0.00000	0.01540	0.0
3	1	2020-05-01	Город	Москва	москва	8611aa017ffffff	55.941586	37.157487	0.00048	0.00054	...	0.10192	0.0	0.00049	NaN	NaN	NaN	NaN	NaN	0.01495	0.0
4	1	2020-05-01	Город	Москва	москва	8611aa637ffffff	55.797494	37.676200	0.00164	0.00179	...	0.09620	0.0	0.00000	0.14444	0.64444	0.01111	0.04444	0.01111	0.01266	0.0

5 rows × 38 columns

Дополнительные данные



Акции и курсы валют - [ИСТОЧНИК](#)

МосБиржа топ

М.видео

Интервал и периодичность

01.01.2020 — 01.08.2022 1 месяц

Имя выходного файла

MVID_200101_220801 .csv

Имя контракта

MVID

Формат

даты ддммгг времени ччммсс

Выдавать время

☐ начала свечи ☒ окончания свечи ☒ московское

Разделитель

полей запятая (,) разрядов нет

Формат записи в файл

TICKER, PER, DATE, TIME, OPEN, HIGH, LOW, CLOSE, VOL

Добавить заголовок файла

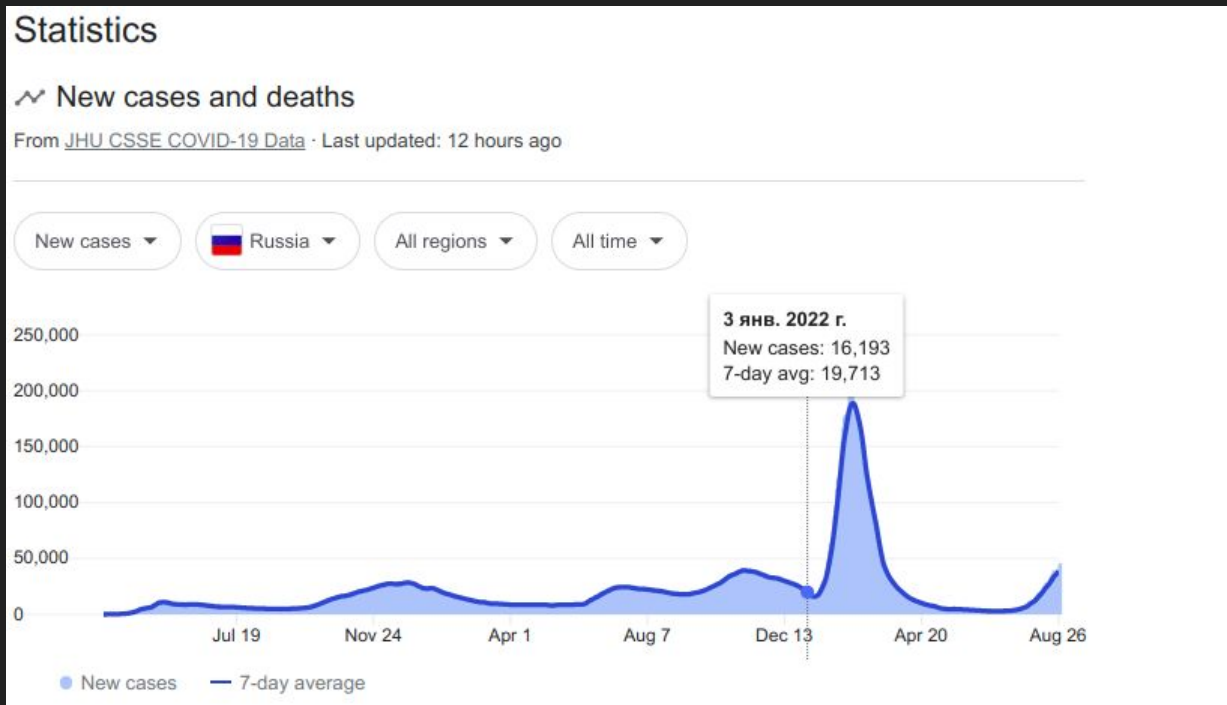
☒

Заполнять периоды без сделок

☐

Получить файл







Коронавирус - ИСТОЧНИК



Информация по городам

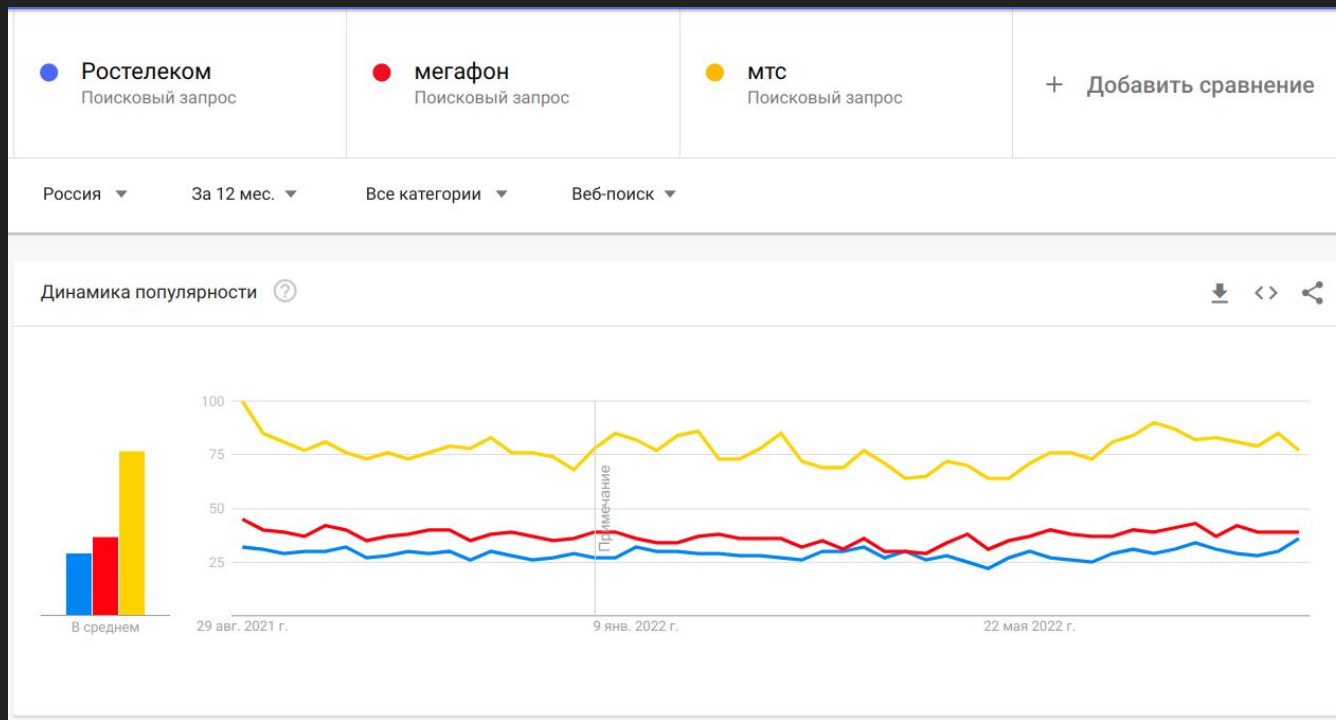
Место	Город	Население на 1 января 2022 года (тыс.) ^[5]	Прирост относительно переписи-2010	Население, тыс. чел. (официальные окончательные итоги переписи 2010 года ^[6])
1	Москва	12 635	9,8 %	11 504
2	Санкт-Петербург	5 378	10,2 %	4 880
3	Новосибирск	1 621	10,0 %	1 474
4	Екатеринбург	1 494	10,7 %	1 350
5	Казань	1 259	10,0 %	1 144
6	Нижний Новгород	1 234	-1,36 %	1 251
7	Челябинск	1 179	4,3 %	1 130
8	Самара	1 137	-2,4 %	1 165
9	Ростов-на-Дону	1 135	4,2 %	1 089
10	Уфа	1 135	6,9 %	1 062
11	Омск	1 126	-3,4 %	1 166
12	Красноярск	1 103	12,9 %	977
13	Воронеж	1 049	7,6 % ^[7]	890 ^[8]
14	Пермь	1 043	5,2 %	991
15	Волгоград	1 001	-1,96 %	1 021

Города миллионники - [источник](#)

3	 Салехард ЯМАЛО-НЕНЕЦКИЙ АВТОНОМНЫЙ ОКРУГ	106 400 ₽
4	 Южно-Сахалинск САХАЛИНСКАЯ ОБЛАСТЬ	99 000 ₽
5	 Магадан МАГАДАНСКАЯ ОБЛАСТЬ	95 200 ₽
6	 Петропавловск-Камчатский КАМЧАТСКИЙ КРАЙ	93 600 ₽
7	 Ханты-Мансийск ХАНТЫ-МАНСИЙСКИЙ АВТОНОМНЫЙ ОКРУГ - ЮГРА	88 300 ₽
8	 Сургут ХАНТЫ-МАНСИЙСКИЙ АВТОНОМНЫЙ ОКРУГ - ЮГРА	83 400 ₽

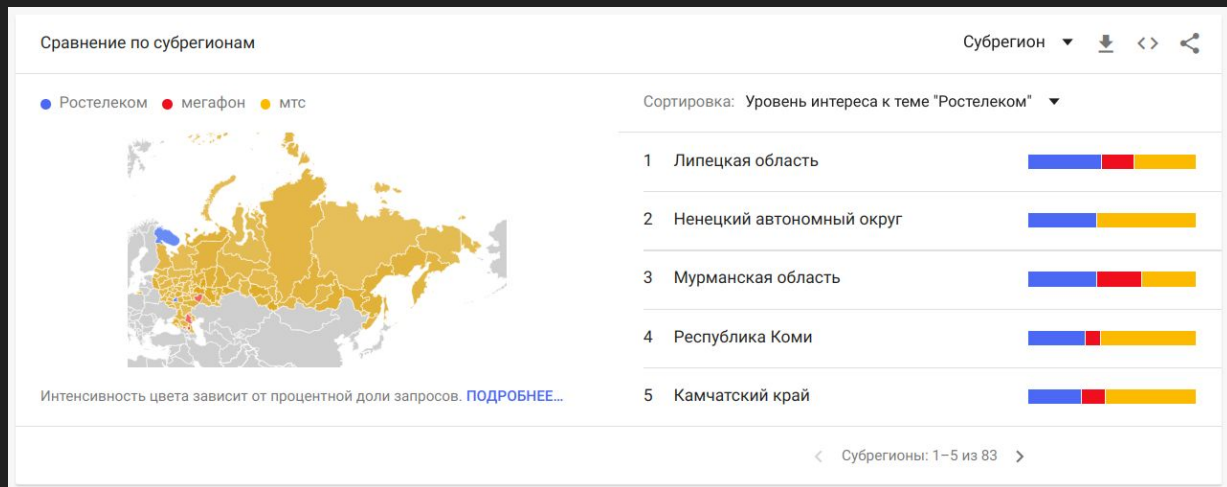
Зарплата по городам - [источник](#)

Google Trends - [ИСТОЧНИК](#)



Google Trends - ИСТОЧНИК

- Спросы на игры
- Жалобы на плохой интернет
- Анализа запросов по конкурентам
- Динамика интереса к Ростелекому



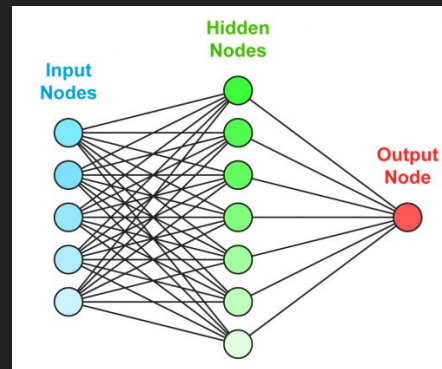
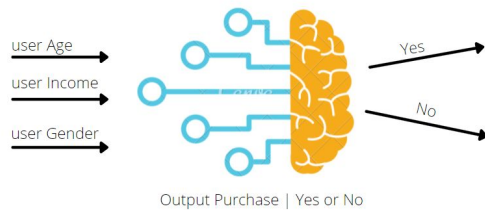
Выжимаем максимум

- Отдельную модель на каждый город
- Постпроцессинг по топу предсказаний внутри каждого города
- Удаляем города, где кол-во покупок ≤ 2
- Подбор гиперпараметров + балансировка классов
- Отбор признаков

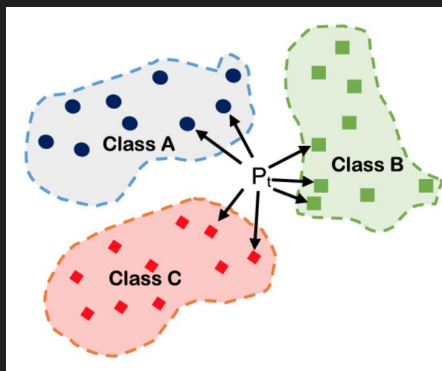
Final Precision - **0.124**

Почему catboost?

Logistic Regression



LightGBM



Использованные инструменты



Не использовано импортное ПО

Масштабируемость

- высокая скорость обучения ($\sim 1m$)
- высокая скорость инференса ($\sim 60ms$ на X_{val})
- быстрый парсинг - либо скачиваются .csv, либо парсим html
- понятный код и легкая воспроизводимость



Спасибо за внимание



Наш гитхаб - [ссылка](#)