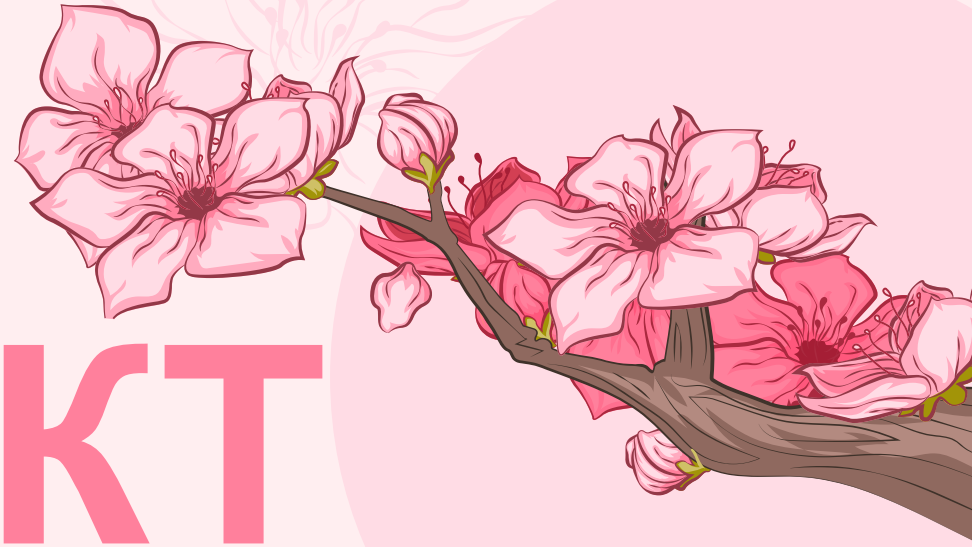
КОМАНДНЫЙ ПРОЕКТ



Выполнили: Воронцова Марина, Ерманова Валерия,
Колесникова Анастасия, Сорокина Анна

Этап 1

Сбор команды, выбор датасета, EDA, baseline

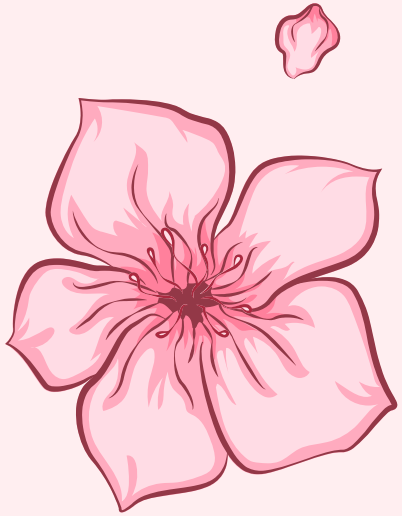
Описание датасета

01 О ЧЁМ ДАТАСЕТ

Продажи различных товаров в сети магазинов “Favorita” в Эквадоре

02 СОСТАВ

Информация про магазин, товар, даты, рекламы, цена нефти, праздничные дни

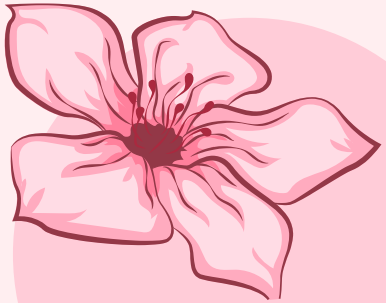


ОБЪЁМ ДАННЫХ 03

Взяли первые 5 миллионов строк (примерно 120 полностью охваченных дней)

ЦЕЛЬ 04

Предсказать продажи, чтобы правильно делать закупки



EDA!

Результаты анализа

1

СТОЛБЕЦ РЕКЛАМЫ

Из 5 млн строк нет ни одной on_promotion

2

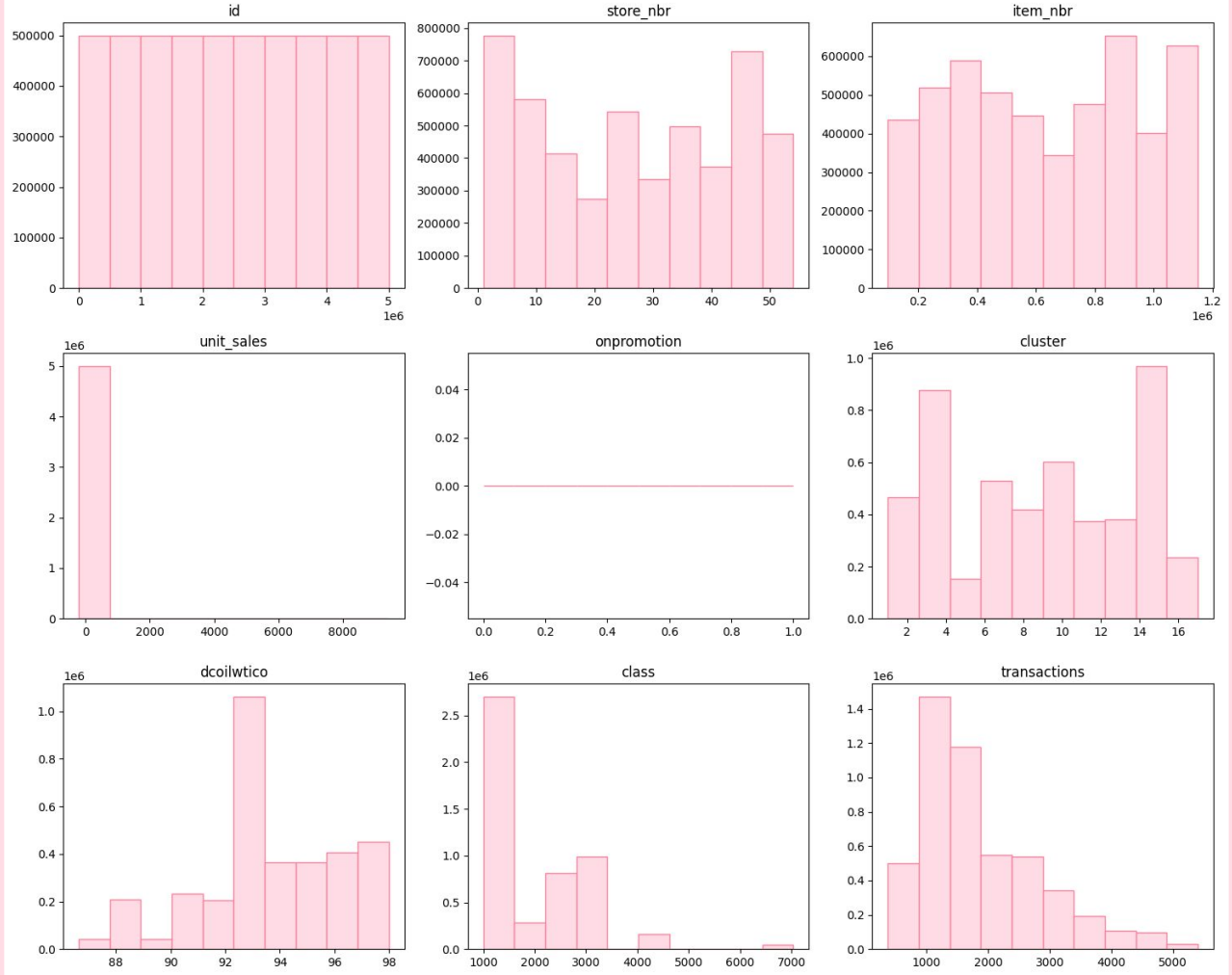
ПРОПУСКИ

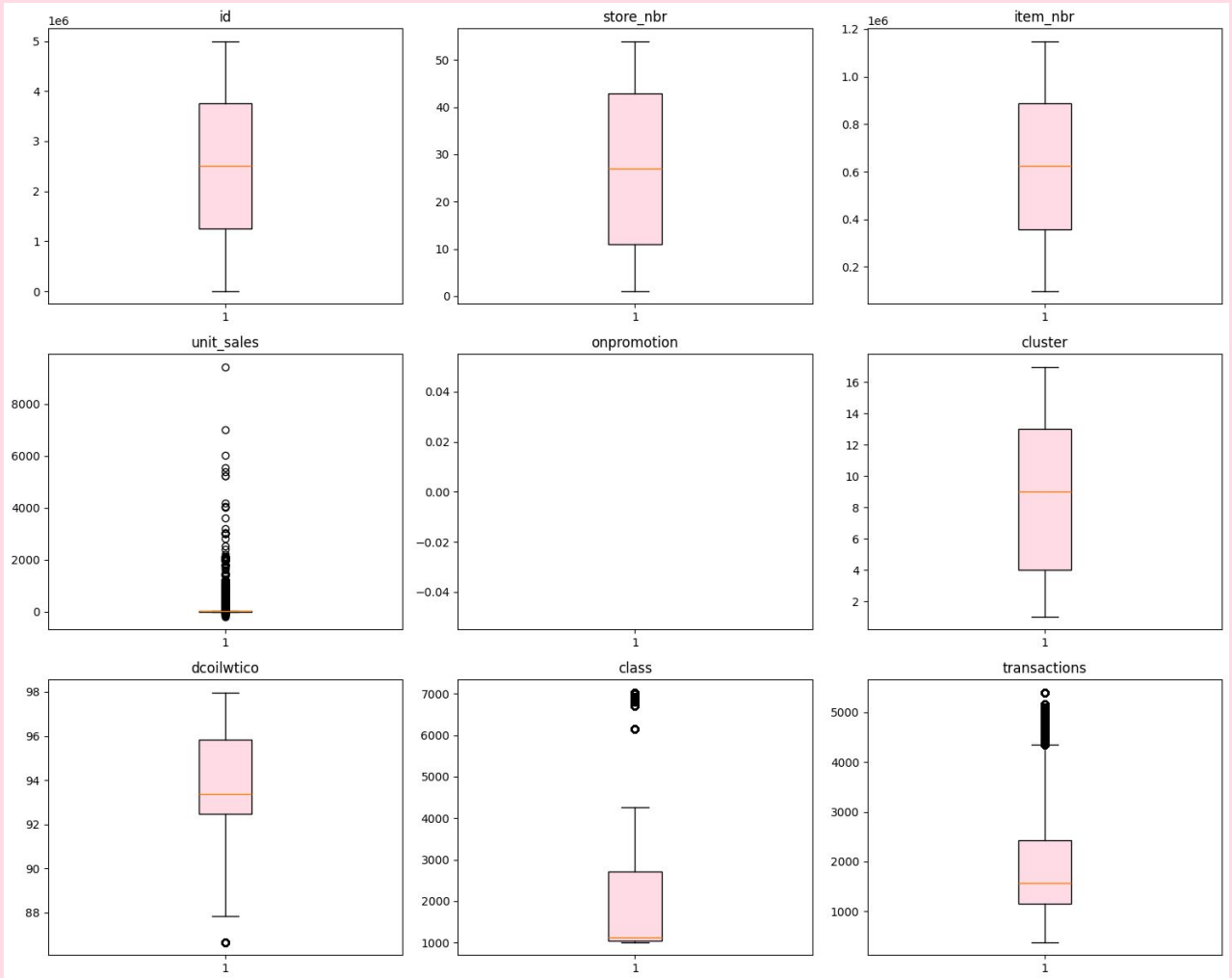
Пропуски есть только в цене нефти и графе праздника в выбранный день

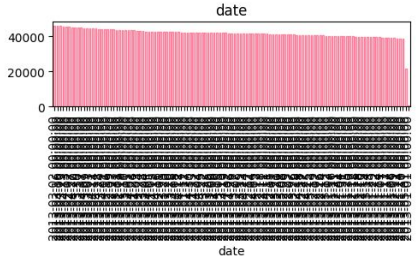
3

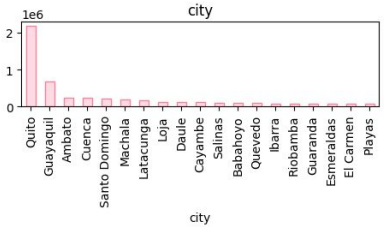
КАТЕГОРИАЛЬНЫЕ ПРИЗНАКИ

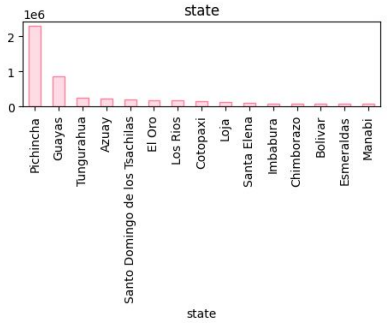
Большинство признаков в данных категориальные

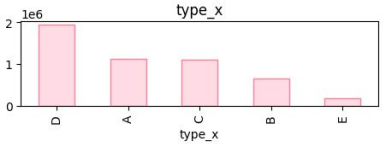

id
store_nbr
item_nbr
unit_sales
onpromotion
cluster
dcoilwtico
class
transactions
1e6

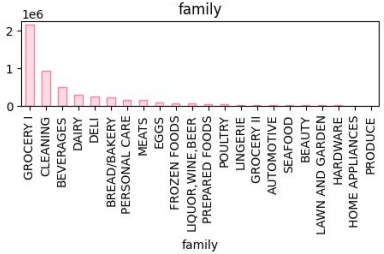

id
1e6
store_nbr
item_nbr
1e6
unit_sales
onpromotion
cluster
dcoilwtico
class
transactions

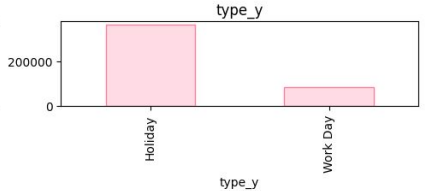

date
40000
20000
0
date

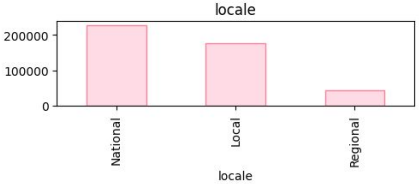

city
1e6
2
1
0
Quito
Guayaquil
Ambato
Cuenca
Santo Domingo
Machala
Latacunga
Loja
Daule
Cayambe
Salinas
Babahoyo
Quevedo
Ibarra
Riobamba
Guaranda
Esmeraldas
El Carmen
Playas
city

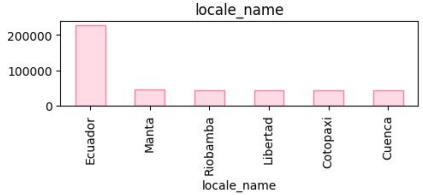

state
1e6
2
1
0
Pichincha
Guayas
Tungurahua
Azuay
Santo Domingo de los Tsachilas
El Oro
Los Rios
Cotopaxi
Loja
Santa Elena
Imbabura
Chimborazo
Bolivar
Esmeraldas
Manabi
state

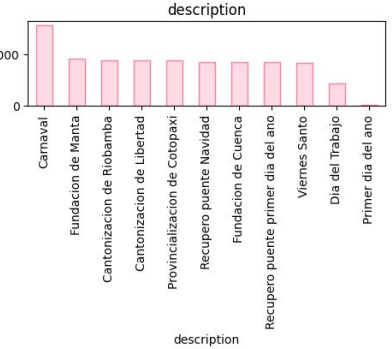

type_x
1e6
2
1
0
D
A
C
B
E
type_x


family
1e6
2
1
0
GROCERY I
CLEANING
BEVERAGES
DAIRY
DELI
BREAD/BAKERY
PERSONAL CARE
MEATS
EGGS
FROZEN FOODS
LIQUOR,WINE,BEER
PREPARED FOODS
POULTRY
LINGERIE
GROCERY II
AUTOMOTIVE
SEAFOOD
BEAUTY
LAWN AND GARDEN
HARDWARE
HOME APPLIANCES
PRODUCE
family


type_y
200000
0
Holiday
Work Day
type_y


locale
200000
100000
0
National
Local
Regional
locale


locale_name
200000
100000
0
Ecuador
Manta
Riobamba
Libertad
Cotopaxi
Cuenca
locale_name

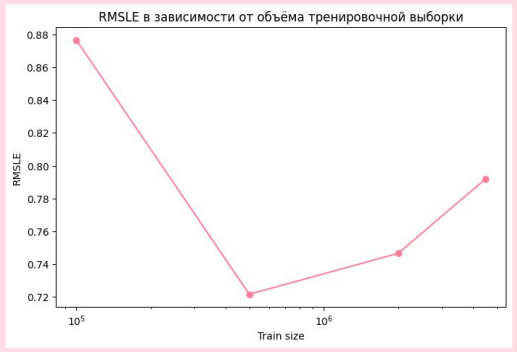

description
50000
0
Carnaval
Fundacion de Manta
Cantonizacion de Riobamba
Cantonizacion de Libertad
Provincializacion de Cotopaxi
Recupero puente Navidad
Fundacion de Cuenca
Recupero puente primer dia del ano
Viernes Santo
Dia del Trabajo
Primer dia del ano
description


Корреляционная матрица
id 1.00 0.00 0.03 -0.00 0.00 -0.45 -0.01 0.00 -0.09 0.01
store_nbr 0.00 1.00 0.01 0.06 -0.00 0.01 0.00 -0.00 0.01 0.24
item_nbr 0.03 0.01 1.00 -0.01 0.00 -0.01 0.05 -0.02 -0.00 0.01
unit_sales -0.00 0.06 -0.01 1.00 0.03 -0.00 -0.05 -0.03 -0.02 0.18
onpromotion
cluster 0.00 -0.00 0.00 0.03 1.00 -0.00 -0.00 -0.01 -0.00 0.20
dcoilwtico -0.45 0.01 -0.01 -0.00 -0.00 1.00 0.00 -0.00 -0.13 0.01
class -0.01 0.00 0.05 -0.05 -0.00 0.00 1.00 -0.28 -0.00 0.02
perishable 0.00 -0.00 -0.02 -0.03 -0.01 -0.00 -0.28 1.00 -0.00 -0.03
transferred -0.09 0.01 -0.00 -0.02 -0.00 -0.13 -0.00 -0.00 1.00 -0.03
transactions 0.01 0.24 0.01 0.18 0.20 0.01 0.02 -0.03 -0.03 1.00
id
store_nbr
item_nbr
unit_sales
onpromotion
cluster
dcoilwtico
class
perishable
transferred
transactions
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

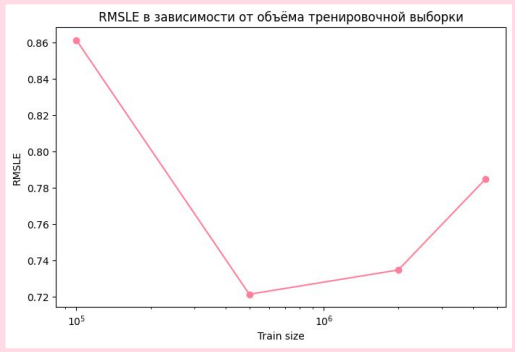
BASELINE!

Метрика – RMSLE (разная цена одинаковых ошибок), модель – lightgbm (скорость, работа с категориальными переменными)

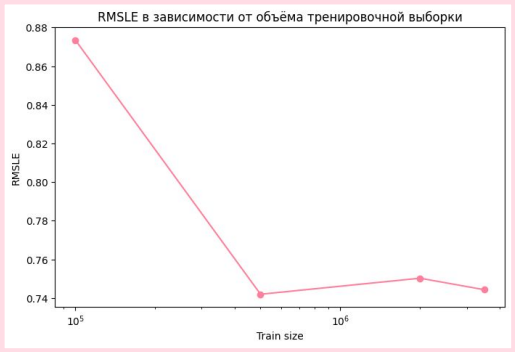
BASELINE



Предсказание на 2 недели



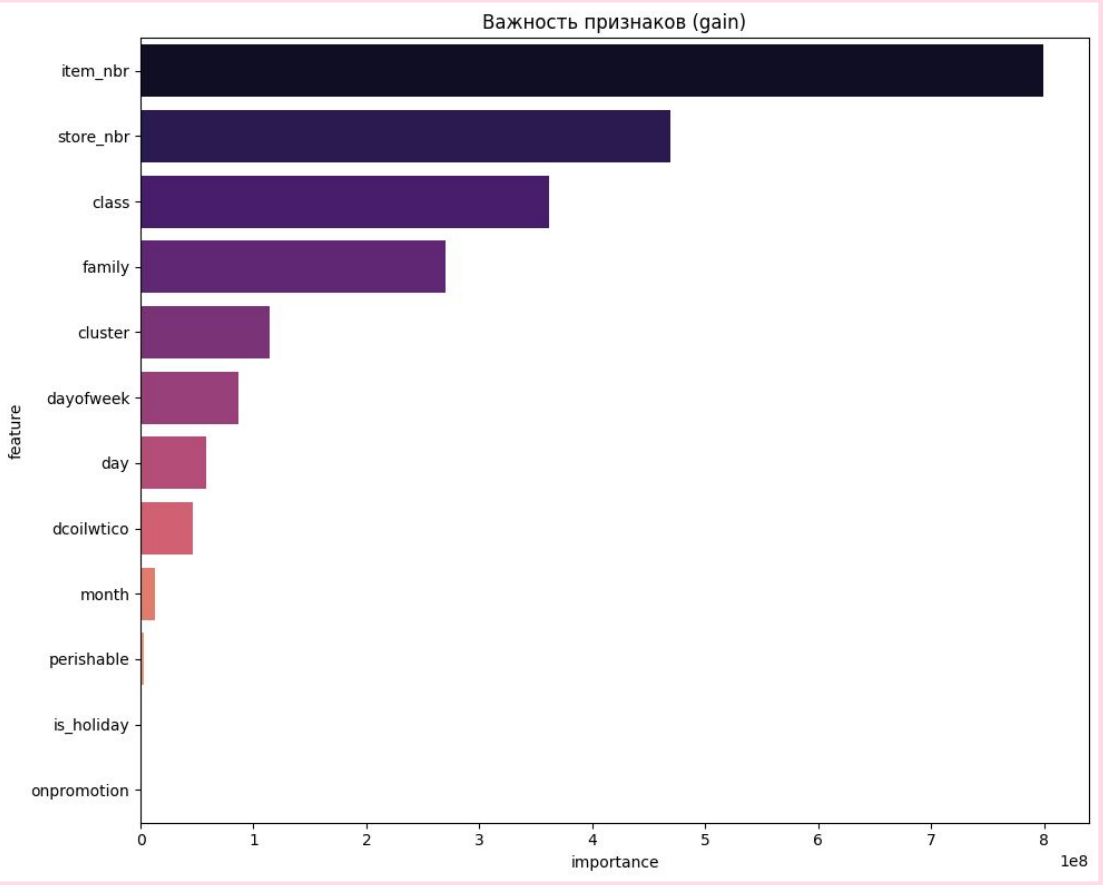
Предсказание на 1 неделю



Предсказание на 4 недели

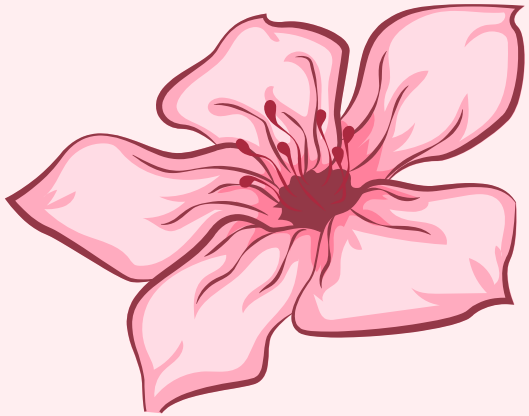
Возможен взлёт RMSLE на слишком больших данных, поэтому берём 2-3 миллиона строк

Важность признаков

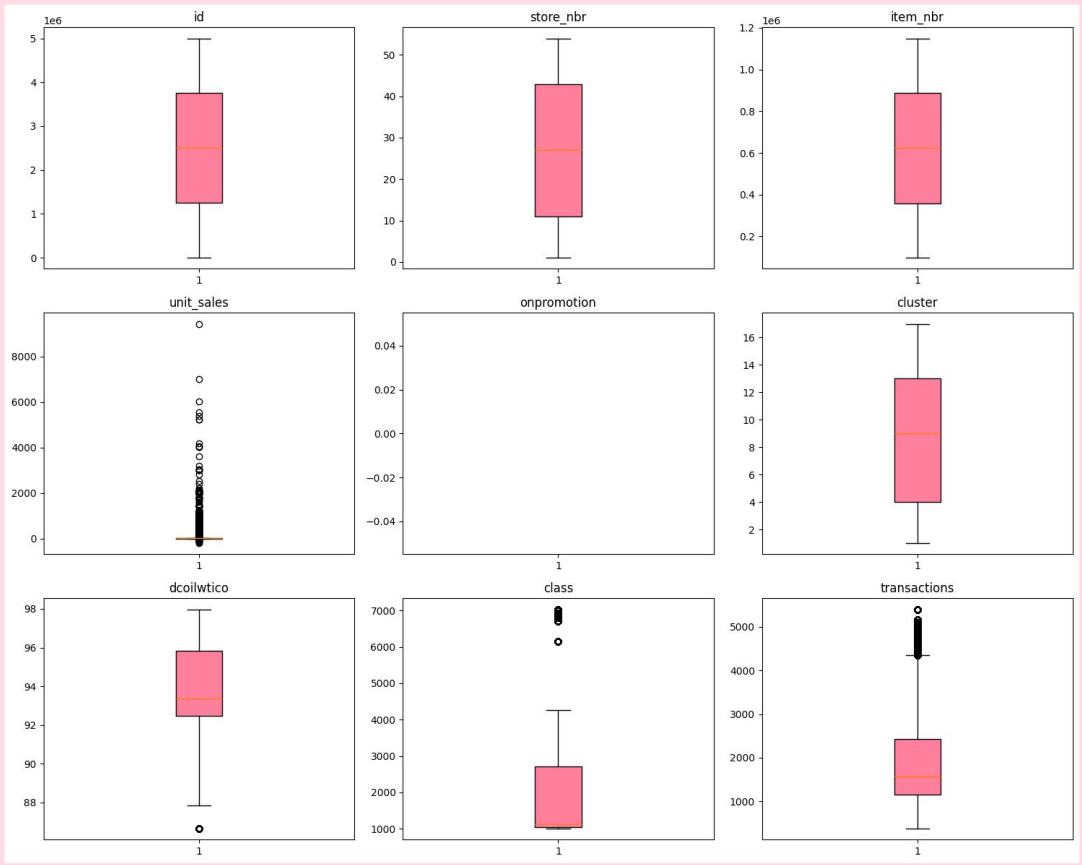


ЭТАП 2

Работа с аномалиями и генерация признаков



Аномалии и выбросы



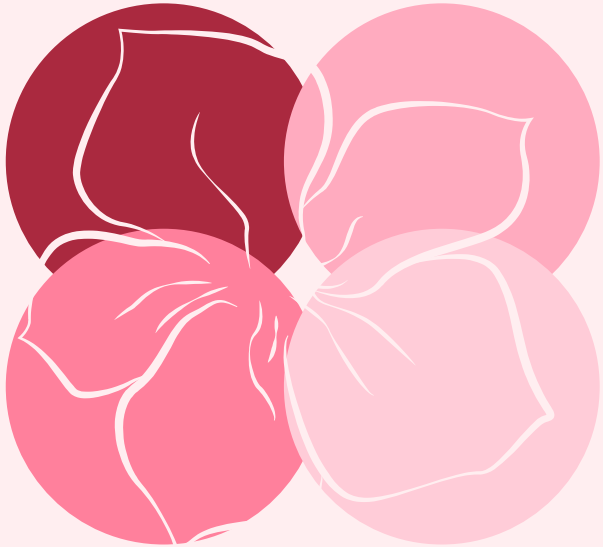
Работа с выбросами

01 ТАРГЕТ

"Выбросов" в таргете целая десятая, но вообще это не выбросы а важная информация для модели – некоторые товары продаются больше других.

02 ЦЕНА НЕФТИ

Выбросов в цене нефти очень мало, большая часть сосредоточена в двух средних квартилях (больше 99%). Такое количество выбросов можно игнорировать (и они даже не слишком сильно выбиваются, всего лишь 86 вместо 87)



03 СЕМЕЙСТВА

Особой корреляции между редкостью семейства и редкостью класса нет, также и с выручкой. Есть один класс который заметно выше остальных по продажам (около 6 раз), есть также другие с высокими продажами, но их мы отловим на обработке больших продаж. То есть с редкими категориями и семействами нам ничего не надо делать

ML для поиска сложных выбросов

```
date: 100000 аномалий
city: 0 аномалий
state: 0 аномалий
type_x: 0 аномалий
family: 4403 аномалий
type_y: 0 аномалий
locale: 0 аномалий
locale_name: 0 аномалий
description: 14 аномалий
```

В категориях аналогично, аномалии в класс и family, но мы их либо отловили в unit_sales, если они важные, либо они не влияют на продажи и не важные. Ну и случайно все даты аномальные

Итог: ml методы тут не нужны, аномалии можно отловить статистически

Шаги 2 этапа

Обработка категориальных переменных

item_nbr и store_nbr закодировали с помощью target encoder, type_x, для type_y, state, city выбрали OneHotEncoding

Признаки, основанные на ближайших соседях

Товары похожих категорий (class, family) и магазины похожего кластера ведут себя одинаково по продажам, поэтому они будут нам полезны.

Формулировка контекстных задач

Так как мы прогнозируем продажи, стоит исследовать несколько гипотез:

1. Сезонные пики
2. Дни до зарплаты, т.к. "новые деньги" всегда дают мотивацию купить что-то классное, новое и полезное

Отбор признаков

unit_sales	1.000000
unit_sales_raw	0.999936
item_nbr_te	0.481036
high_unit_sales	0.475117
class_te	0.262239
similar_items_mean_sales	0.262239
transactions	0.193867
store_nbr_te	0.175219
family_te	0.175204
cluster_mean_sales	0.155362
type_x_0	0.129106
state_11	0.094401
city_15	0.085807
store_nbr	0.058691
dayofweek	0.058025

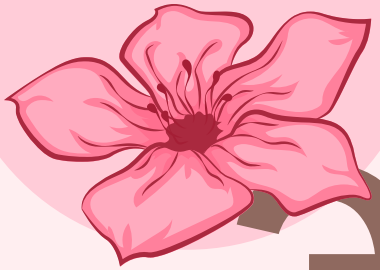
Через фильтры - корреляции,

χ^2, ANOVA;

	feature	importance
18	unit_sales_raw	1499
13	transactions	274
0	id	182
17	day	160
7	class	148
5	dcoilwtico	136
20	item_nbr_te	122
2	item_nbr	76
23	class_te	69
1	store_nbr	64
15	dayofweek	58
6	family	54
21	store_nbr_te	41
4	cluster	40
22	family_te	19
67	cluster_mean_sales	16
27	type_x_3	9
46	state_14	7
8	perishable	7
25	type_x_1	6

Через встроенные методы — L1-регуляризация,

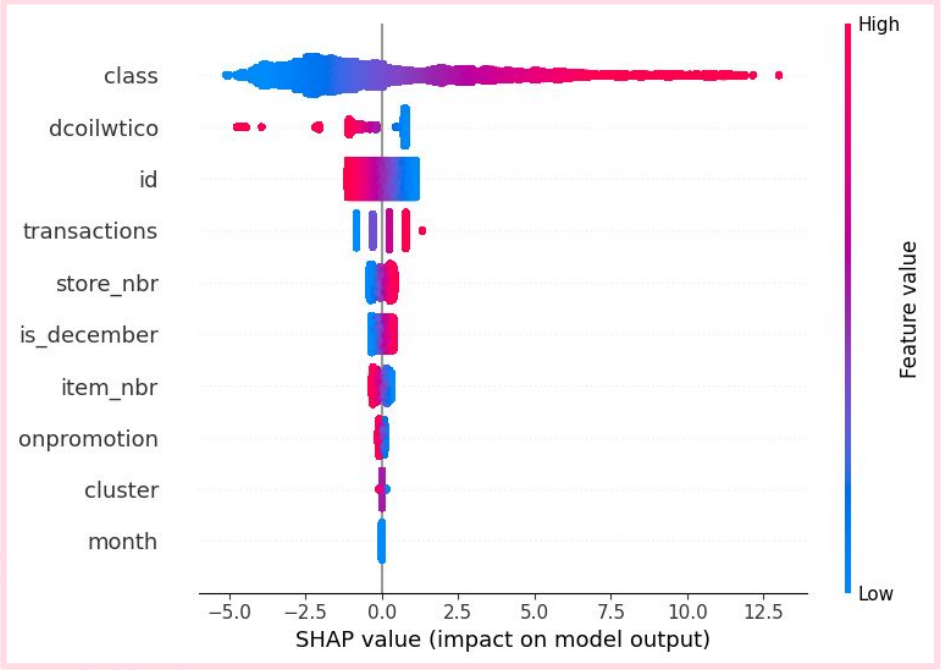
feature importances из деревьев/бустингов



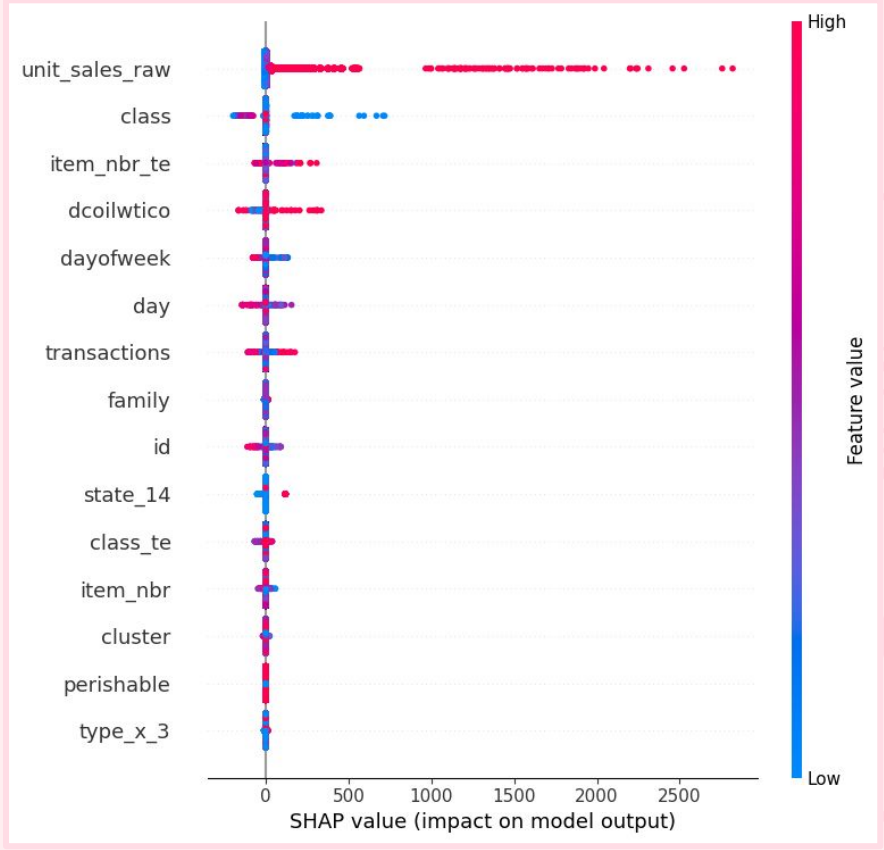
ЭТАП 3

Интерпретация и диагностика моделей

SHAP



Линейная регрессия



LightGBM

SHAP

```
Топ 10 SHAP признаков (линейный случай):
class           2.702551
dcoilwtico      0.836929
id              0.564704
transactions    0.541204
store_nbr       0.231949
is_december     0.193466
item_nbr        0.176668
onpromotion     0.068127
cluster         0.032620
month           0.000000
```

```
Топ 15 SHAP признаков (лгб случай):
unit_sales_raw    7.448394
class             0.013610
item_nbr_te       0.006343
dcoilwtico        0.005016
dayofweek         0.004118
day               0.004055
transactions      0.003862
family            0.003719
id                0.002841
state_14          0.002588
class_te          0.002253
item_nbr          0.001377
cluster           0.000697
perishable        0.000585
type_x_3          0.000465
```

```
Совпадение признаков:
{'cluster', 'id', 'item_nbr', 'transactions', 'class', 'dcoilwtico'}
```

THE SLIDE TITLE GOES HERE!

Do you know what helps you make your point clear? Lists like this one:

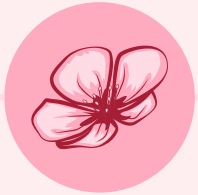
- They’re simple
- You can organize your ideas
- You’ll never forget to buy milk!

The audience won’t miss the point of your presentation

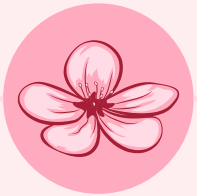
BLOSSOM MIND MAP

M

Mars is a cold place



EARTH



J

Jupiter is a gas giant

BLOSSOM MIND MAP

01

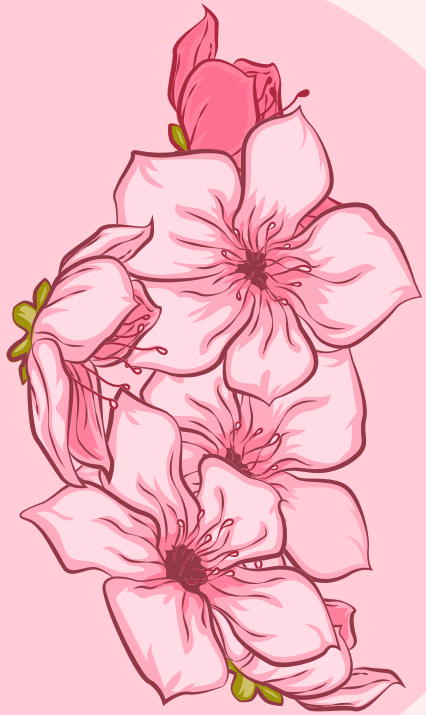
Mercury is the closest planet to the Sun

02

Venus has a beautiful name, but it’s very hot

03

Despite being red, Mars is a cold place

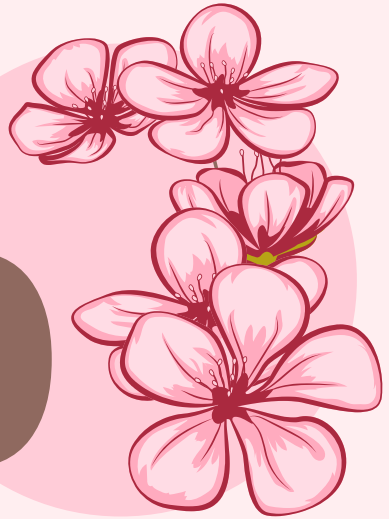


“HERE’S A QUOTE, WORDS FULL OF WISDOM THAT CAN MAKE THE READER GET INSPIRED.”

—Someone Famous

5,000

Big numbers catch your audience’s attention and make ideas understandable



REVIEWING CONCEPTS

MERCURY

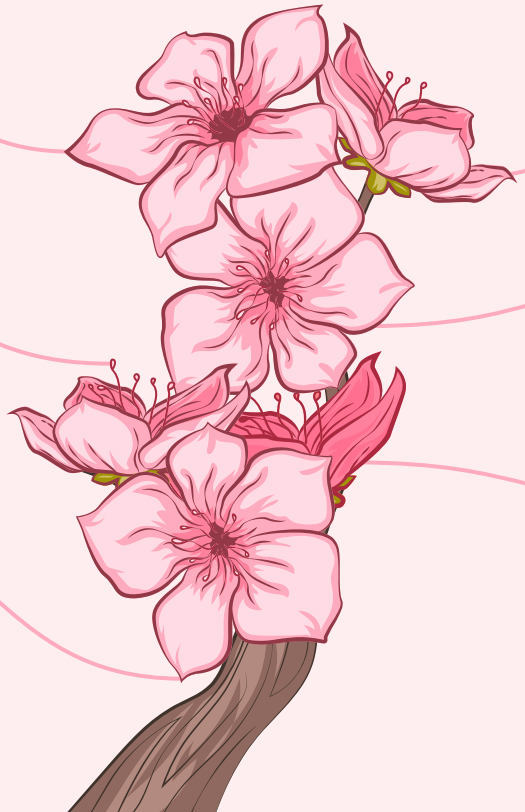
It’s the closest planet to the Sun

JUPITER

It’s the biggest in the Solar System

NEPTUNE

Neptune is the farthest planet



VENUS

Venus has a beautiful name

MARS

Despite being red, Mars is cold

SATURN

It’s a gas giant and has rings

Спасибо за внимание!