ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Проект на курсе ML

Состав команды:

- Белоглазов Михаил Юрьевич
- Воронин Георгий Владимирович
- Карпов Денис Денисович
- Перминов Матвей Максимович

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Этап 1. Датасет и EDA

Датасет

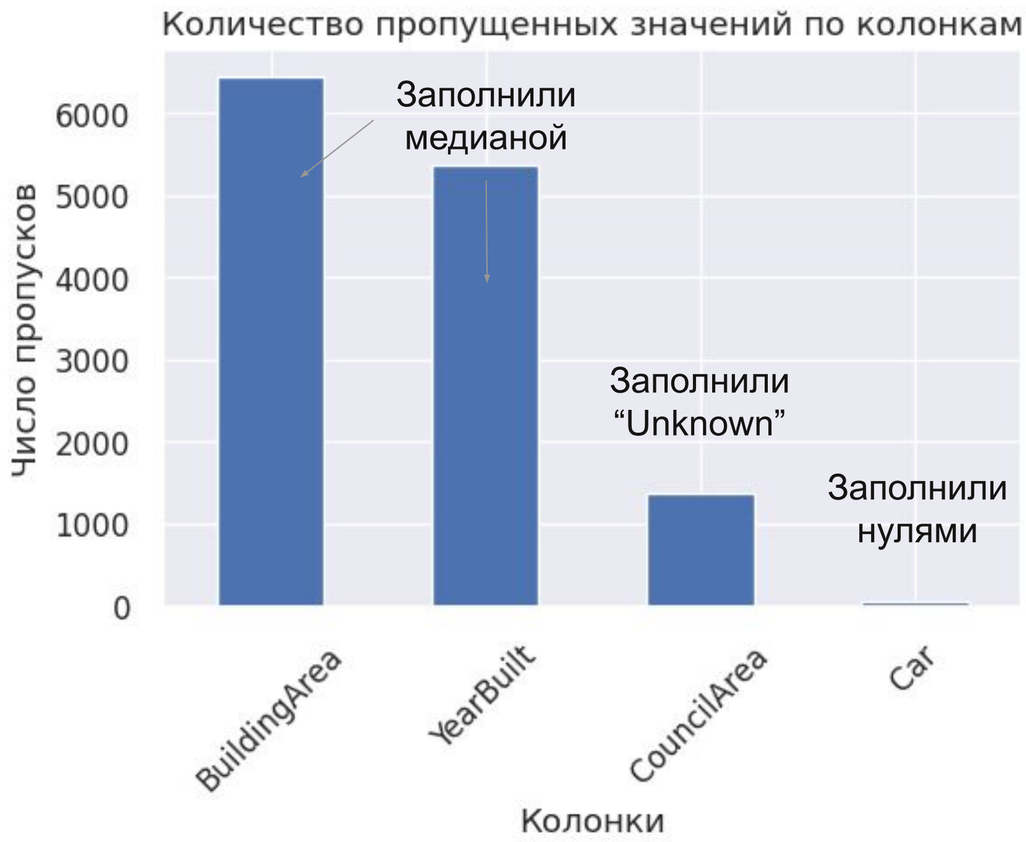
Melbourne Housing - содержит данные о продажах разной недвижимости (квартиры, дома, таунхаусы) на аукционе в Мельбурне (Австралия)

- 13 000 строк

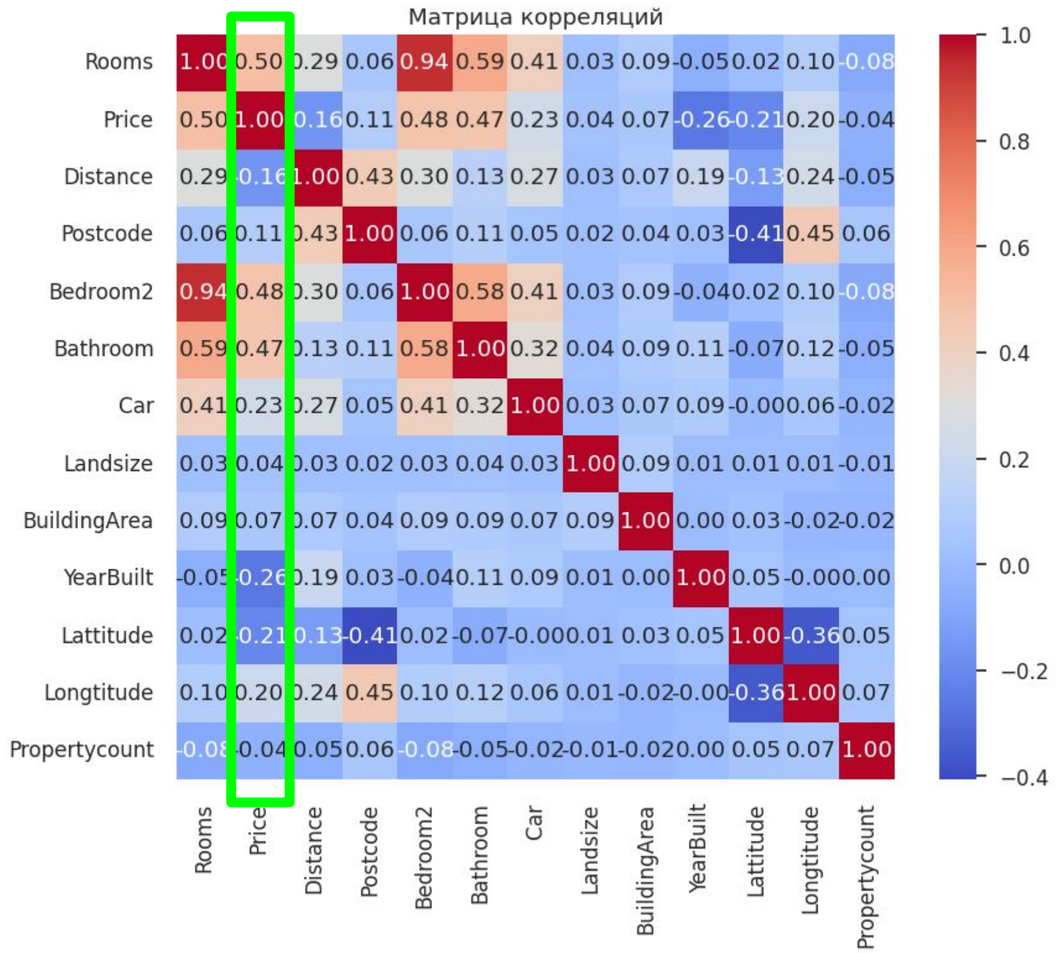
- Целевая переменная: Price (цена, за которую продан дом)

- Есть временные и пространственные признаки

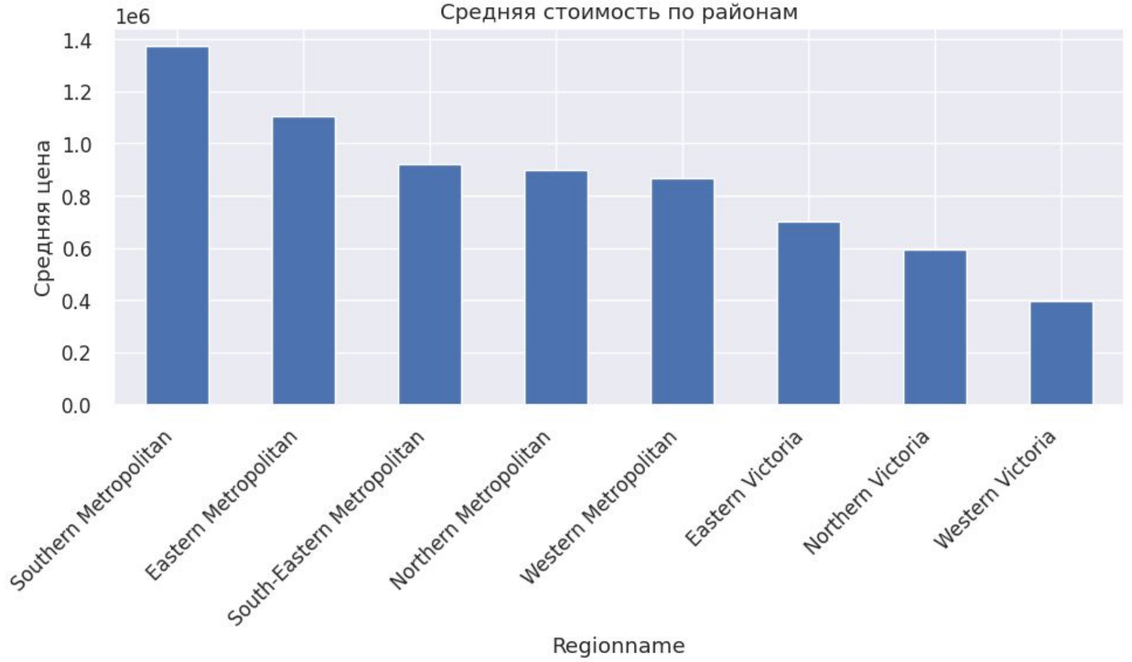
Пропущенные значения



Корреляции




Топ-10 признаков по абсолютной корреляции с target
Price
Rooms
Bedroom2
Bathroom
YearBuilt
Car
Lattitude
Longtitude
Distance
Postcode
BuildingArea
Propertycount
Landsize
1.00
0.50
0.48
0.47
0.26
0.23
0.21
0.20
0.16
0.11
0.07
0.04
0.04
0.0
0.2
0.4
0.6
0.8
1.0
|Корреляция с target|


Средняя стоимость по районам
1e6
Средняя цена
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Southern Metropolitan
Eastern Metropolitan
South-Eastern Metropolitan
Northern Metropolitan
Western Metropolitan
Eastern Victoria
Northern Victoria
Western Victoria
Regionname


1e6
Средняя стоимость по количеству спален
Средняя цена
0.0
0.5
1.0
1.5
7.0
5.0
6.0
20.0
9.0
4.0
8.0
3.0
0.0
10.0
2.0
1.0
Bedroom2


Hume
Moreland
North Eastern
Melbourne
Brimbank
Melton
Outer Eastern Melbourne
Inner
Eastern
Melbourne
Western
Melbourne
MELBOURNE
PORT
PHILLIP
Southern Melbourne
Bayside
Peninsula
BASS STRAIT


Средняя стоимость по Postcode
1e6
2.0
1.5
1.0
0.5
0.0
Средняя цена
отсортированные значения Postcode


Средняя стоимость по PostcodeParted
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Средняя цена
отсортированные значения PostcodeParted


Распределение значений Price
Количество
2000
1000
0
0
2
4
6
8
Price
1e6

Baseline

	rmse_train	rmse_val
Catboost_baseline	203991.248249	269203.054416
Linreg_baseline	253387.388905	379754.136154

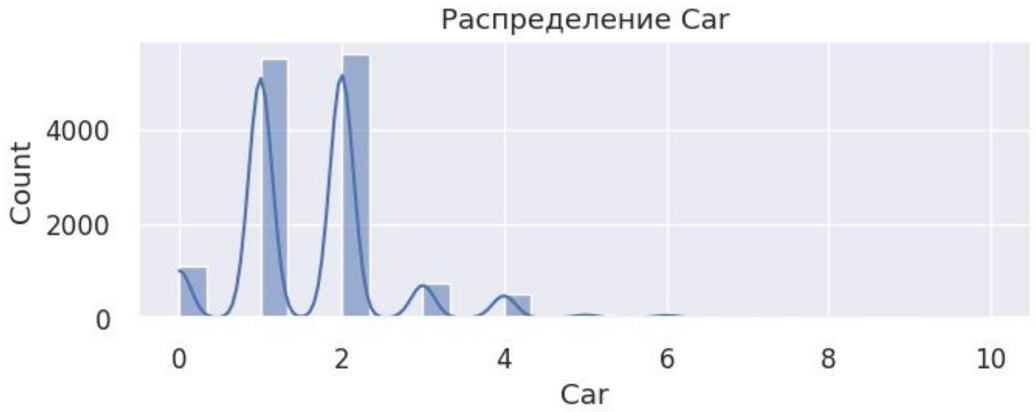
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

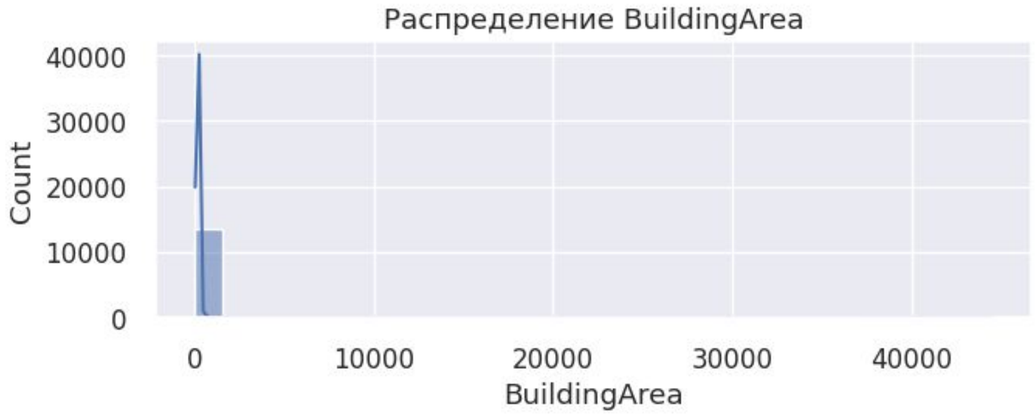
Этап 2. Работа с аномалиями и генерация признаков

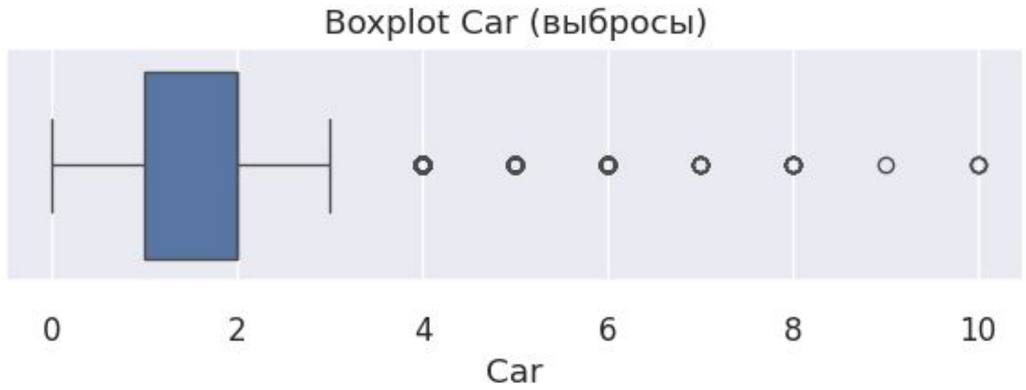
Анализ таргета

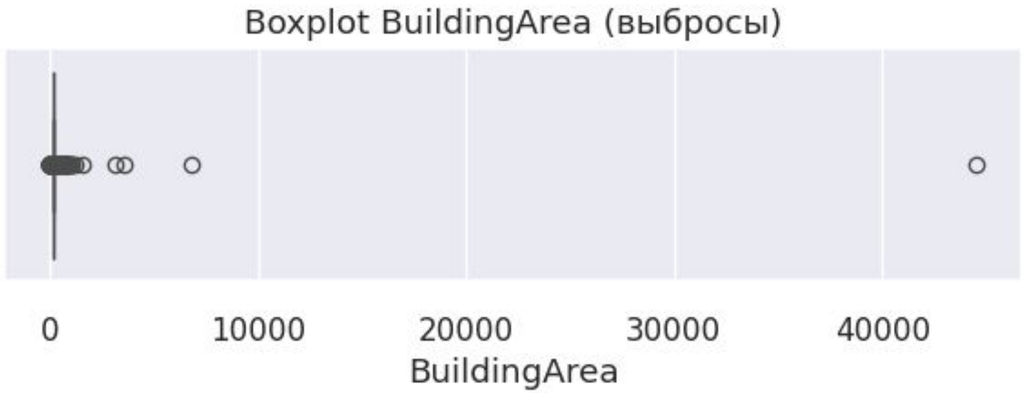

```
=== Выбросы по Z-score (>|3|) ===
Количество выбросов: 232

=== Выбросы по IQR (1.5 * IQR) ===
Q1: 650000.0 Q3: 1330000.0 IQR: 680000.0
Нижняя граница: -370000.0 Верхняя граница: 2350000.0
Количество выбросов: 612
```

Анализ Car и BuildingArea







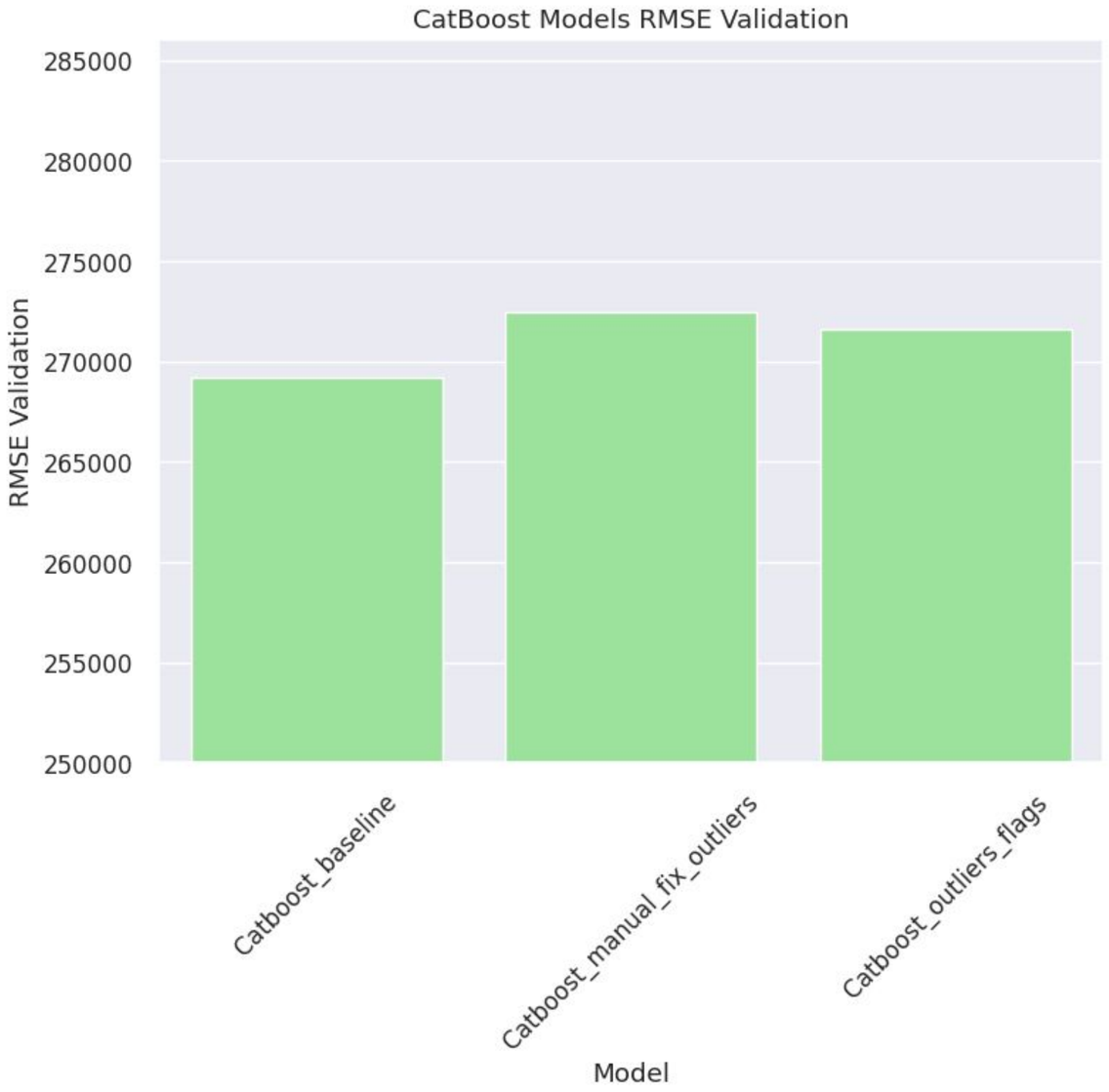


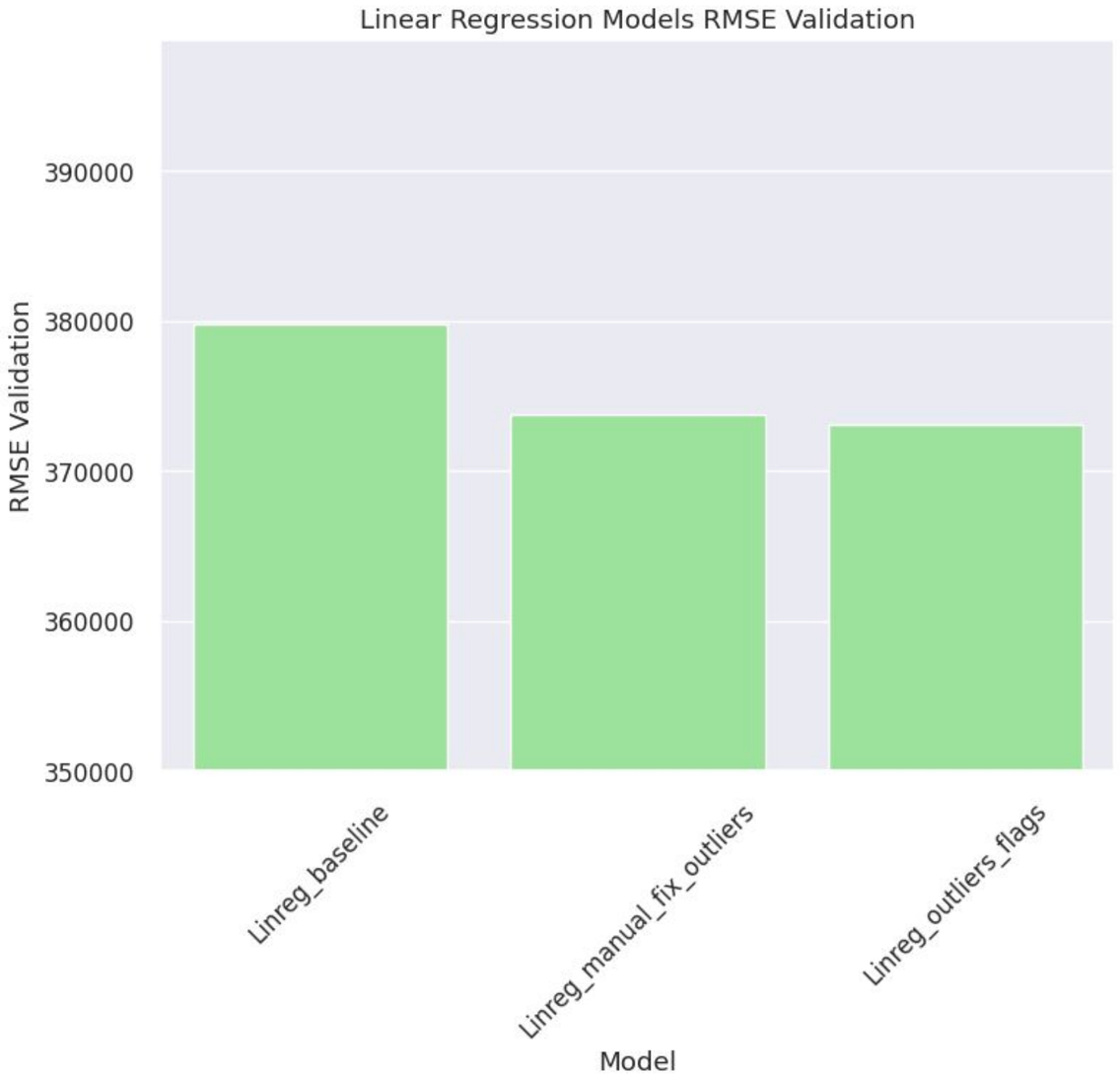
Пример выброса

Поставлю площадь здания 44 515 м^2

(около 6-7 футбольных полей)

Добавим признаки связанные с аномалиями

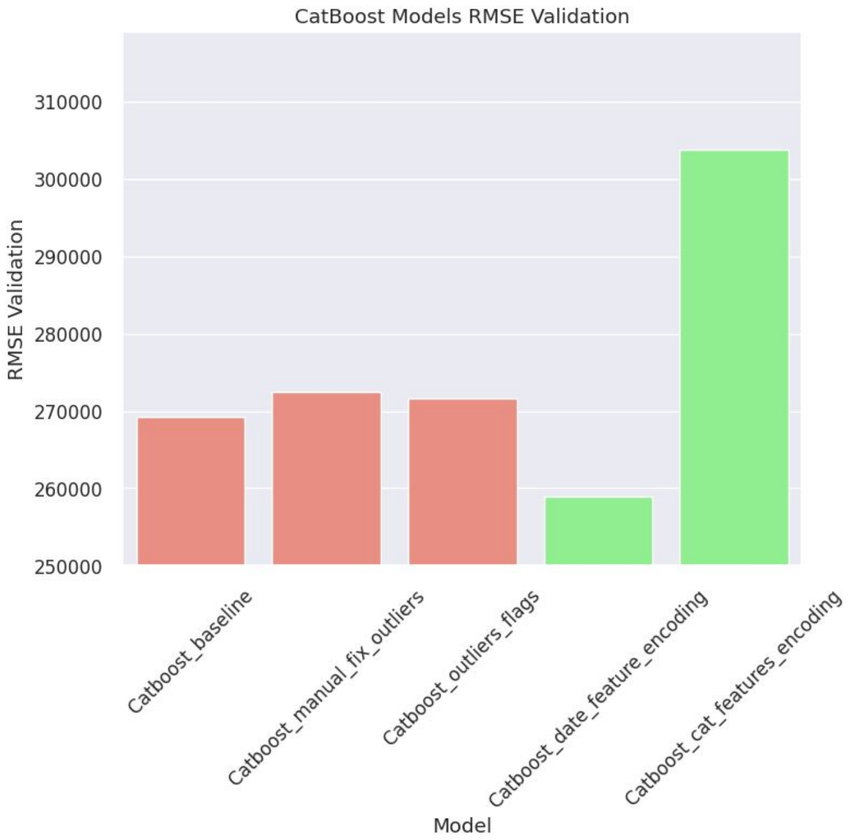


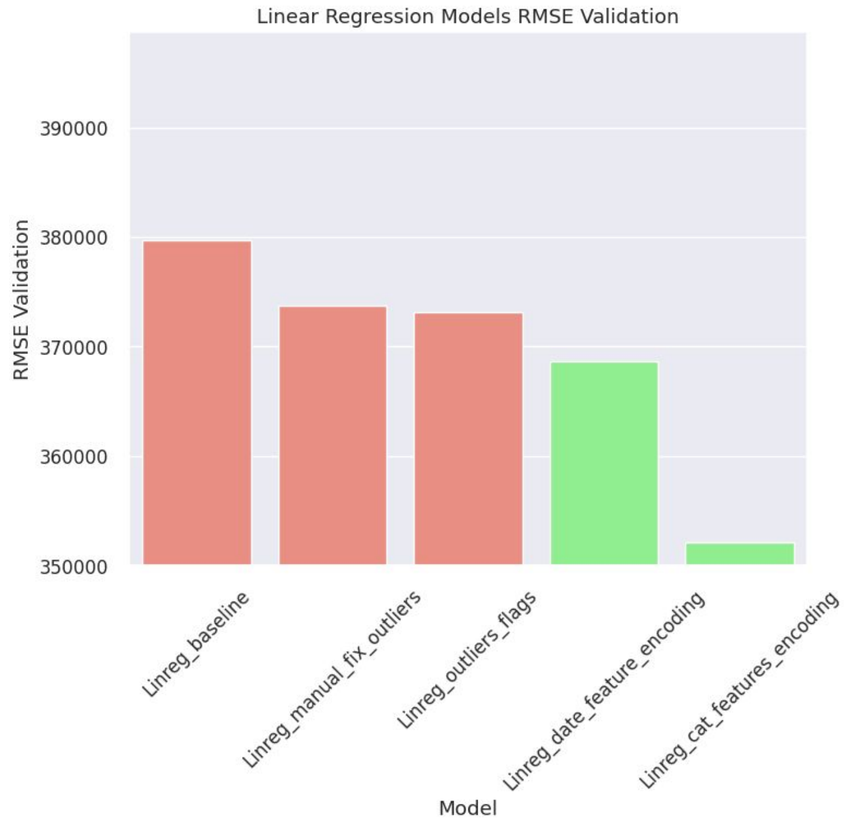


Обработка категориальных признаков

Промежуточные метрики

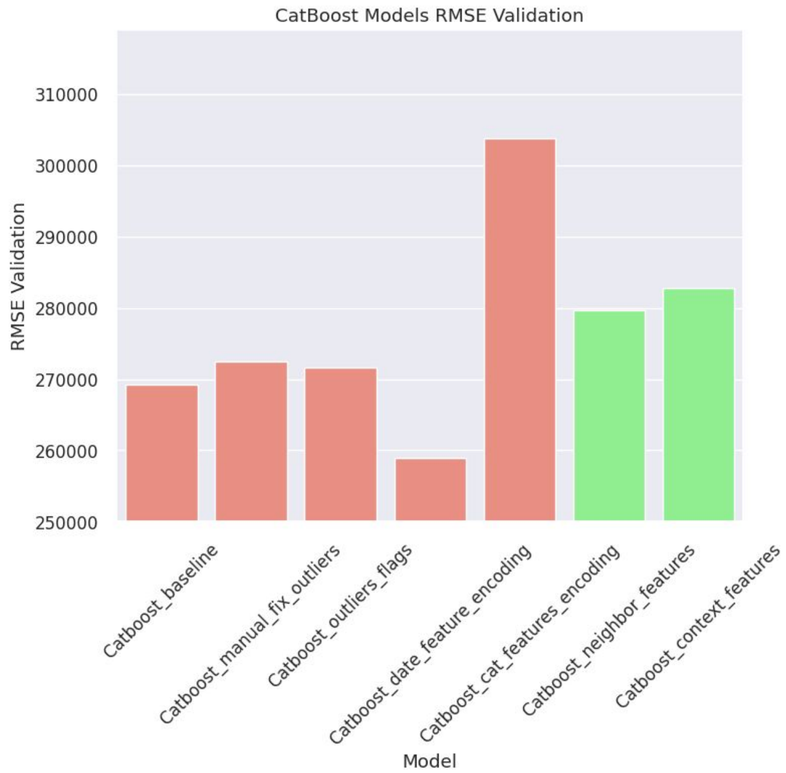
Обработали временной признак Date, разбив его на отдельные части.

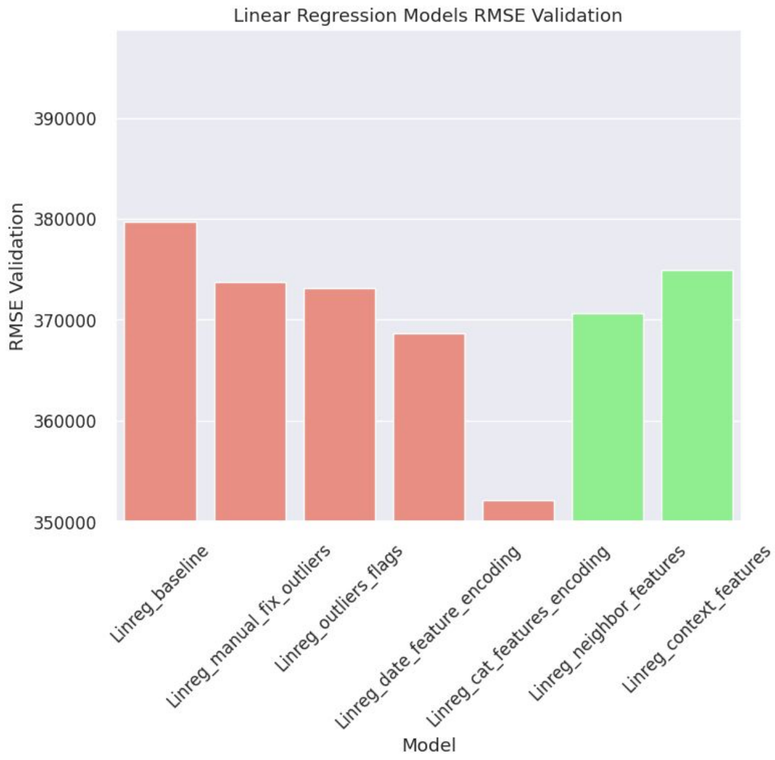




Генерация новых признаков

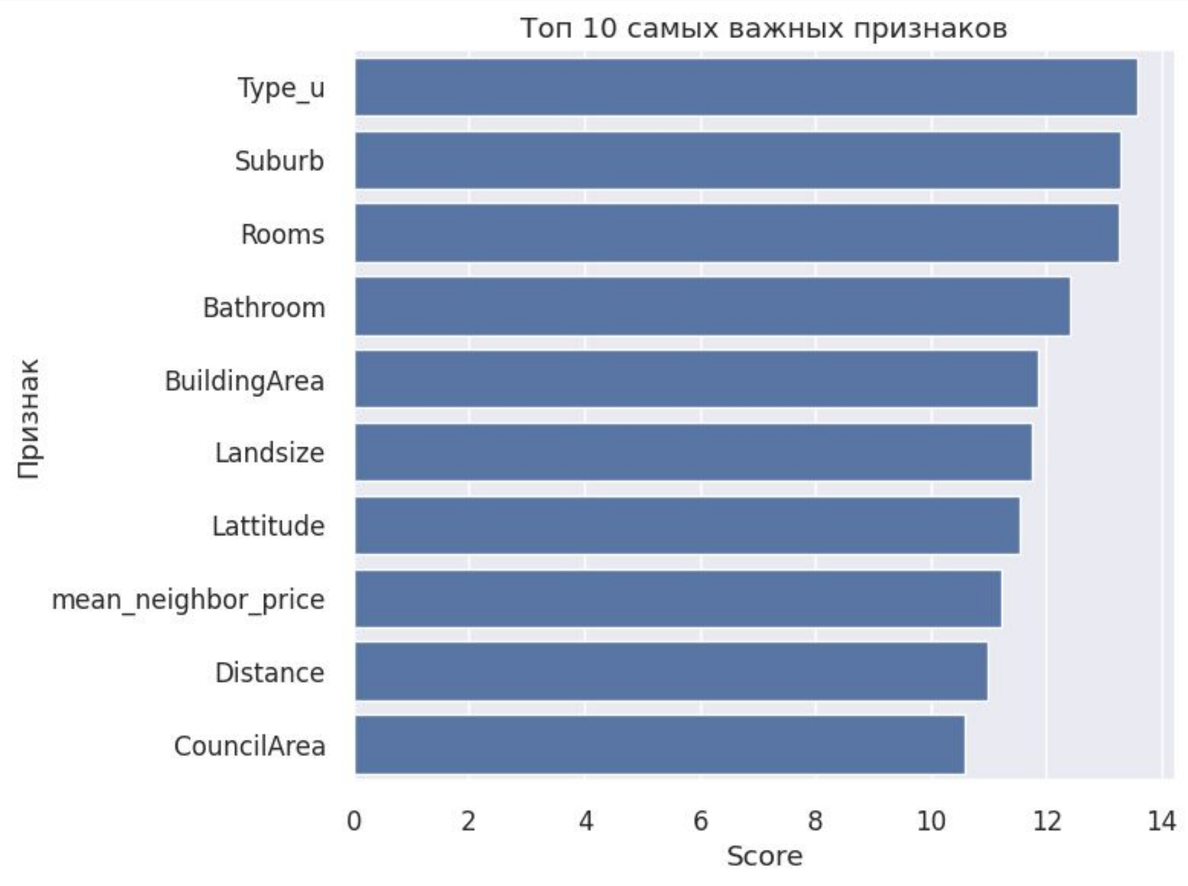
- Признаки, основанные на ближайших географических соседях (средняя цена, расстояние до ближайшего соседа, среднее расстояние до соседей)
- Контекстные признаки:
- Отношение ванных комнат к общему числу комнат
- Отношение жилой площади к площади участка

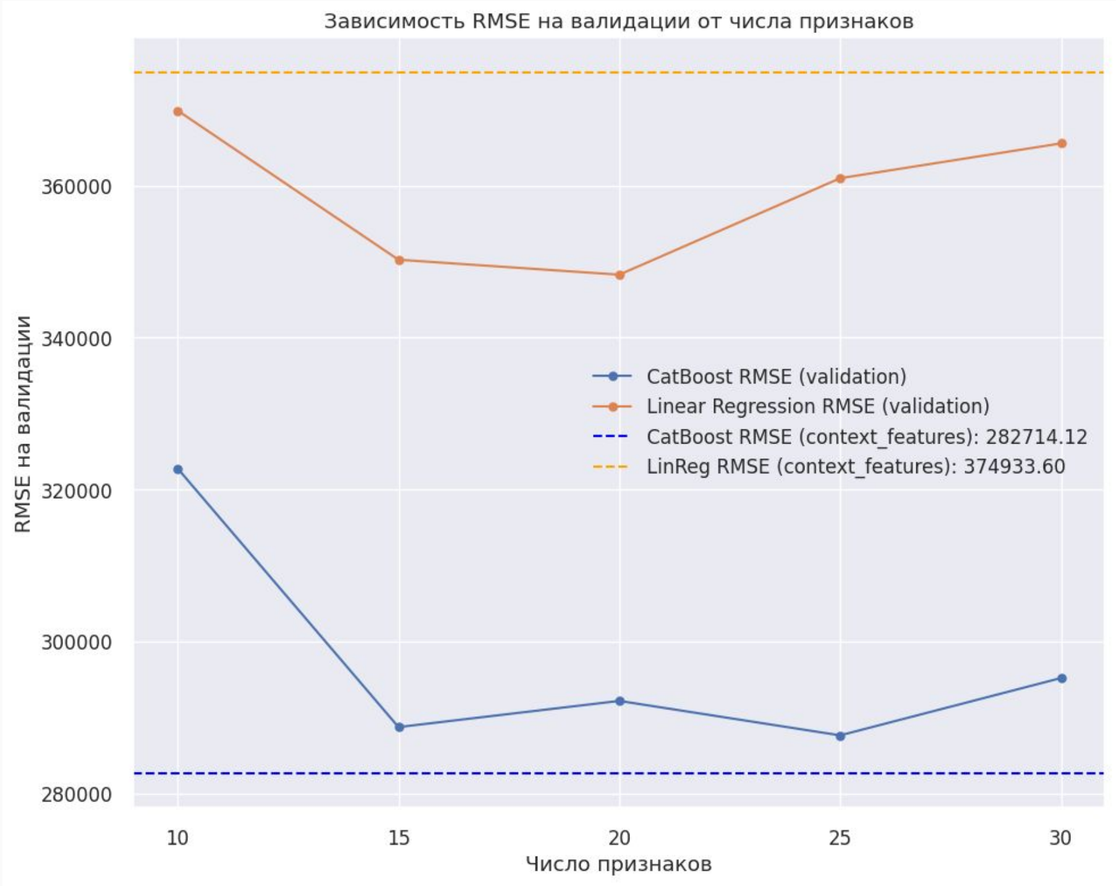




Отбор признаков

Мы использовали комплексную систему отбора признаков, которая учитывает разные способы отбора с весами

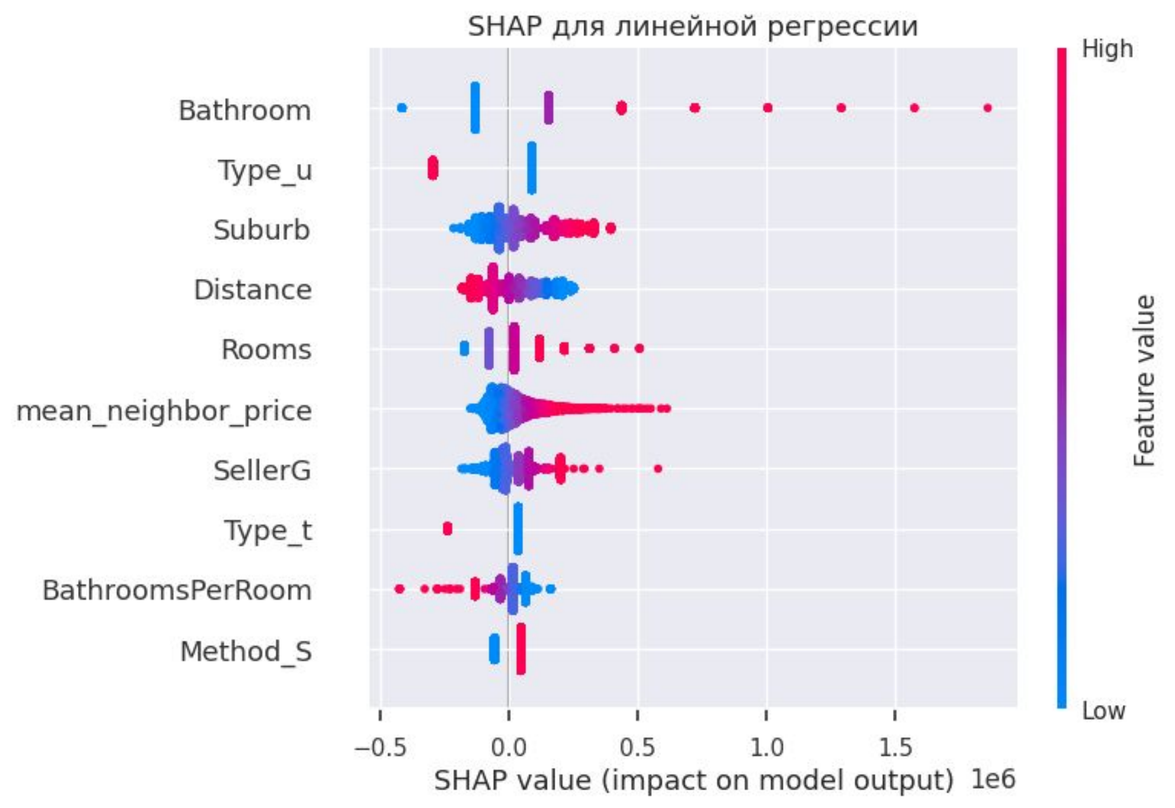


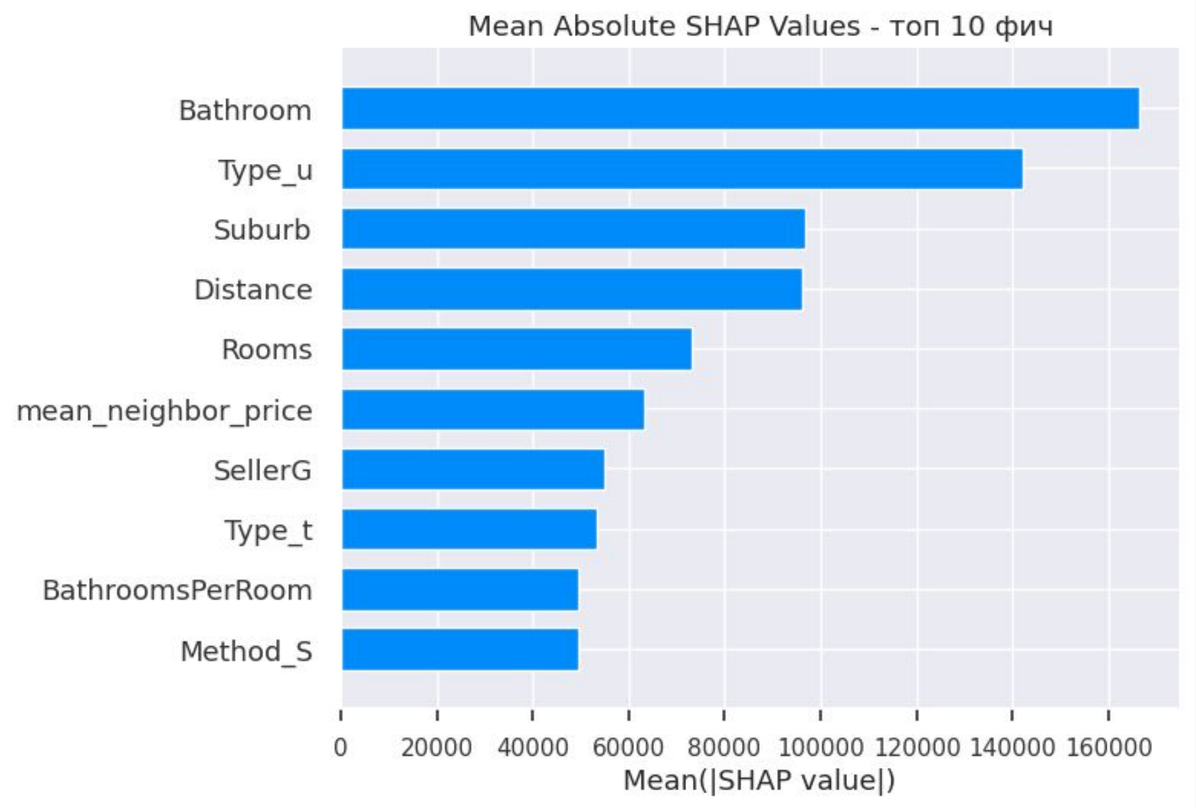


ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

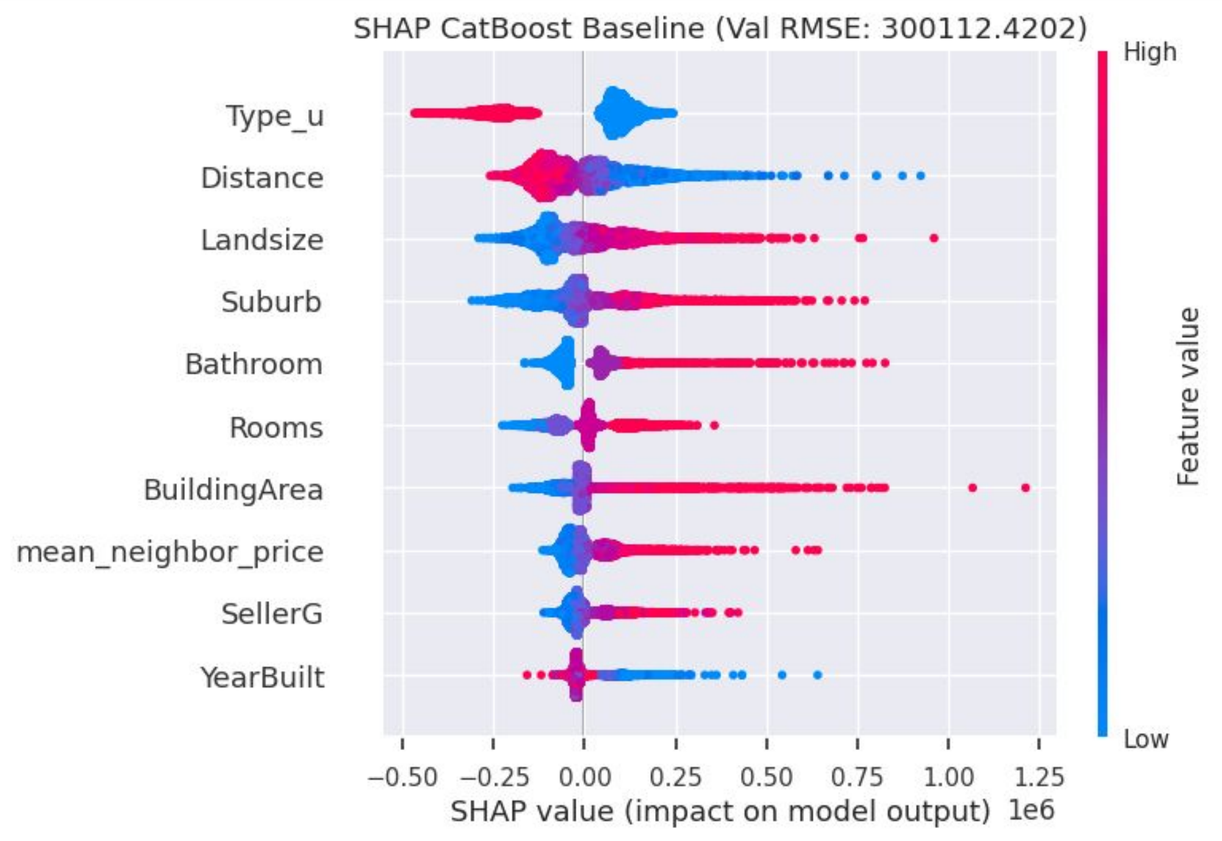
Этап 3. Интерпретация и диагностика моделей

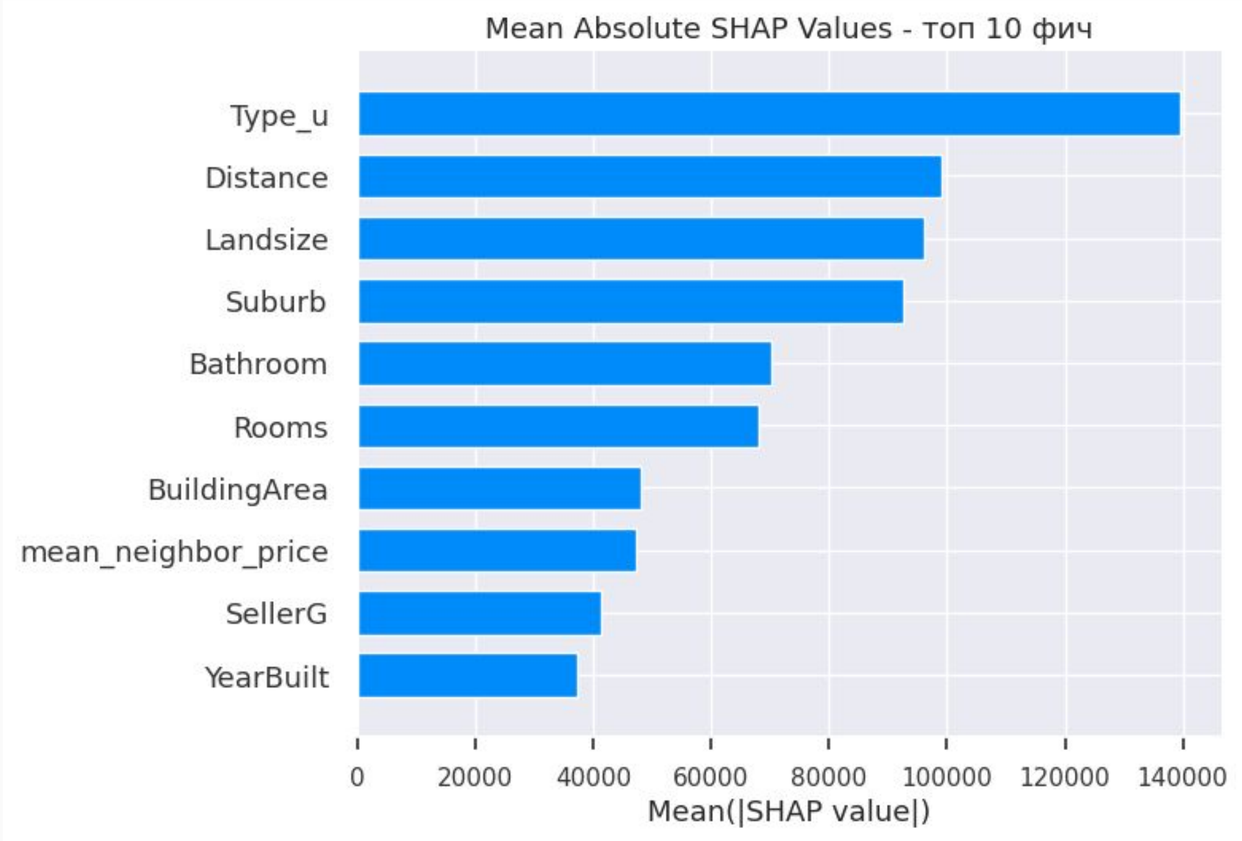
SHAP для линрега



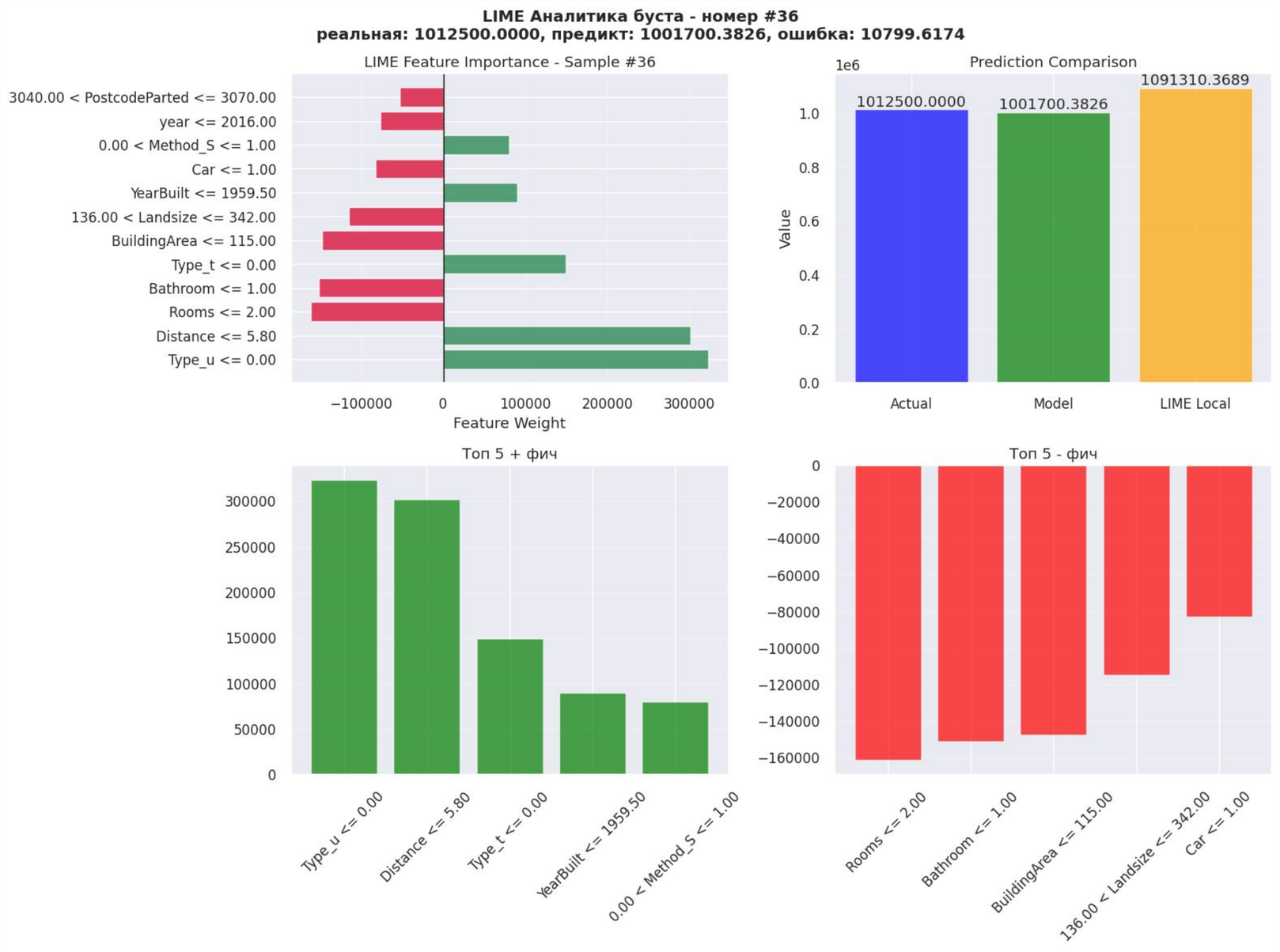


SHAP для кэтбуста

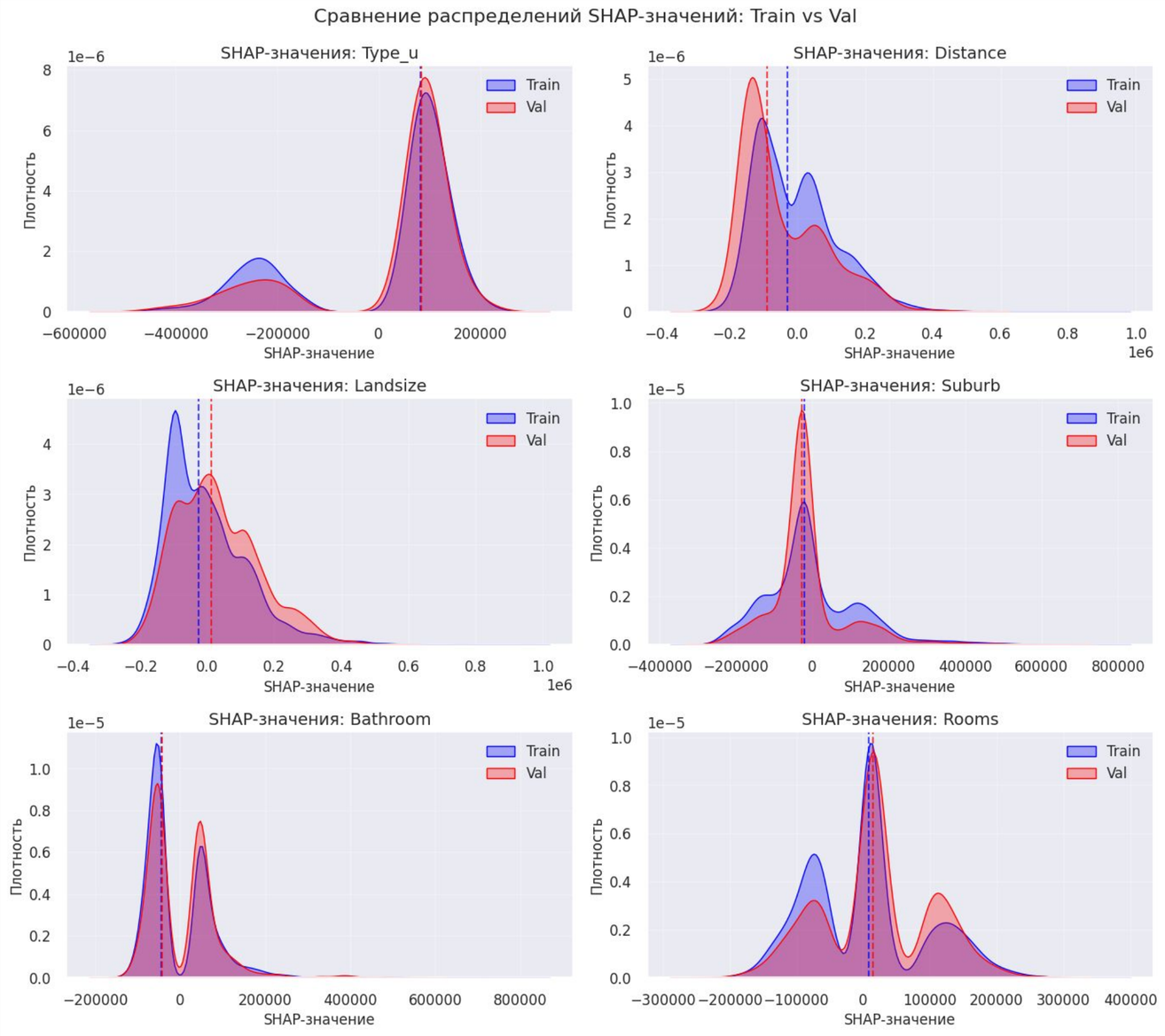




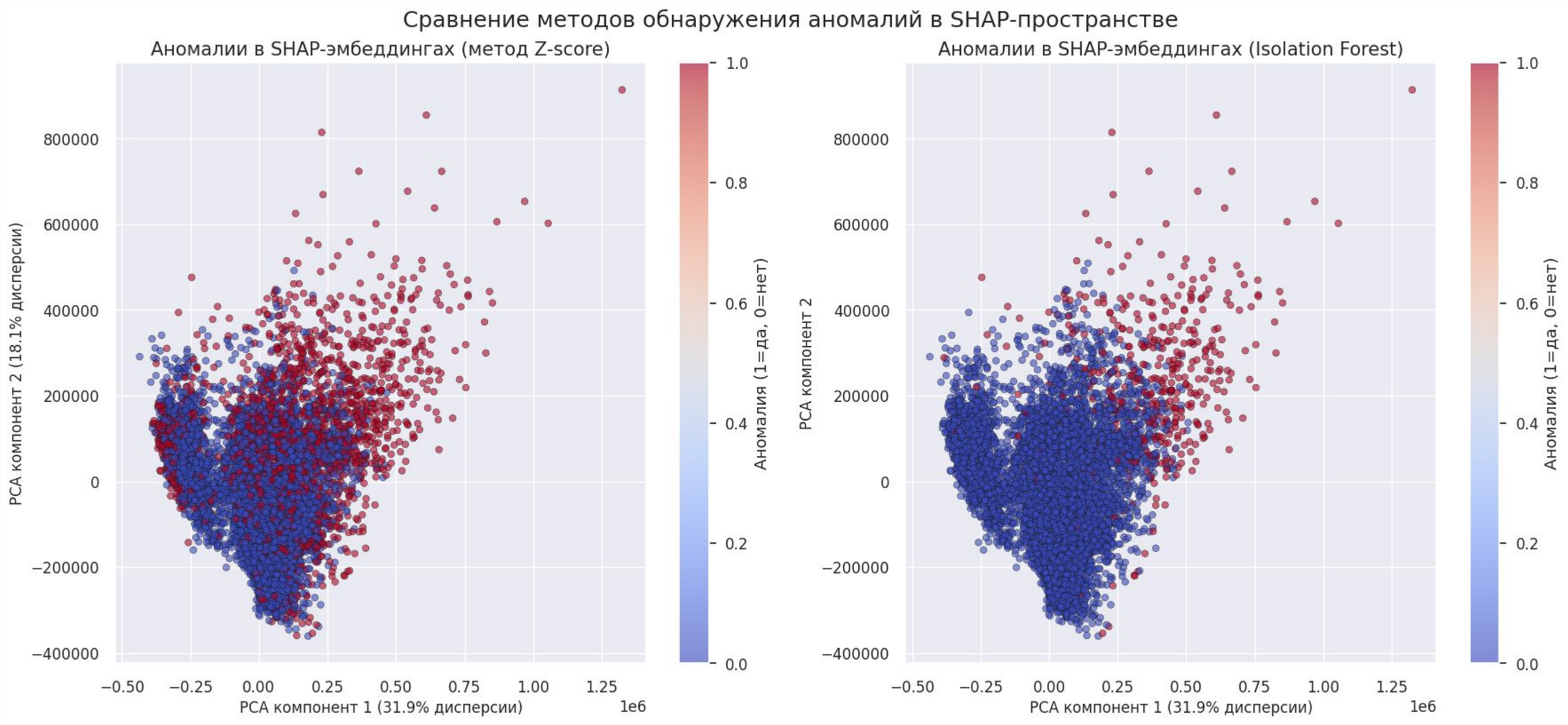
LIME


LIME Аналитика буста - номер #36
реальная: 1012500.0000, предикт: 1001700.3826, ошибка: 10799.6174
LIME Feature Importance - Sample #36
3040.00 < PostcodeParted <= 3070.00
year <= 2016.00
0.00 < Method_S <= 1.00
Car <= 1.00
YearBuilt <= 1959.50
136.00 < Landsize <= 342.00
BuildingArea <= 115.00
Type_t <= 0.00
Bathroom <= 1.00
Rooms <= 2.00
Distance <= 5.80
Type_u <= 0.00
−100000
0
100000
200000
300000
Feature Weight
Prediction Comparison
1e6
1012500.0000
1001700.3826
1091310.3689
Value
Actual
Model
LIME Local
Топ 5 + фич
Type_u <= 0.00
Distance <= 5.80
Type_t <= 0.00
YearBuilt <= 1959.50
0.00 < Method_S <= 1.00
Топ 5 - фич
Rooms <= 2.00
Bathroom <= 1.00
BuildingArea <= 115.00
136.00 < Landsize <= 342.00
Car <= 1.00

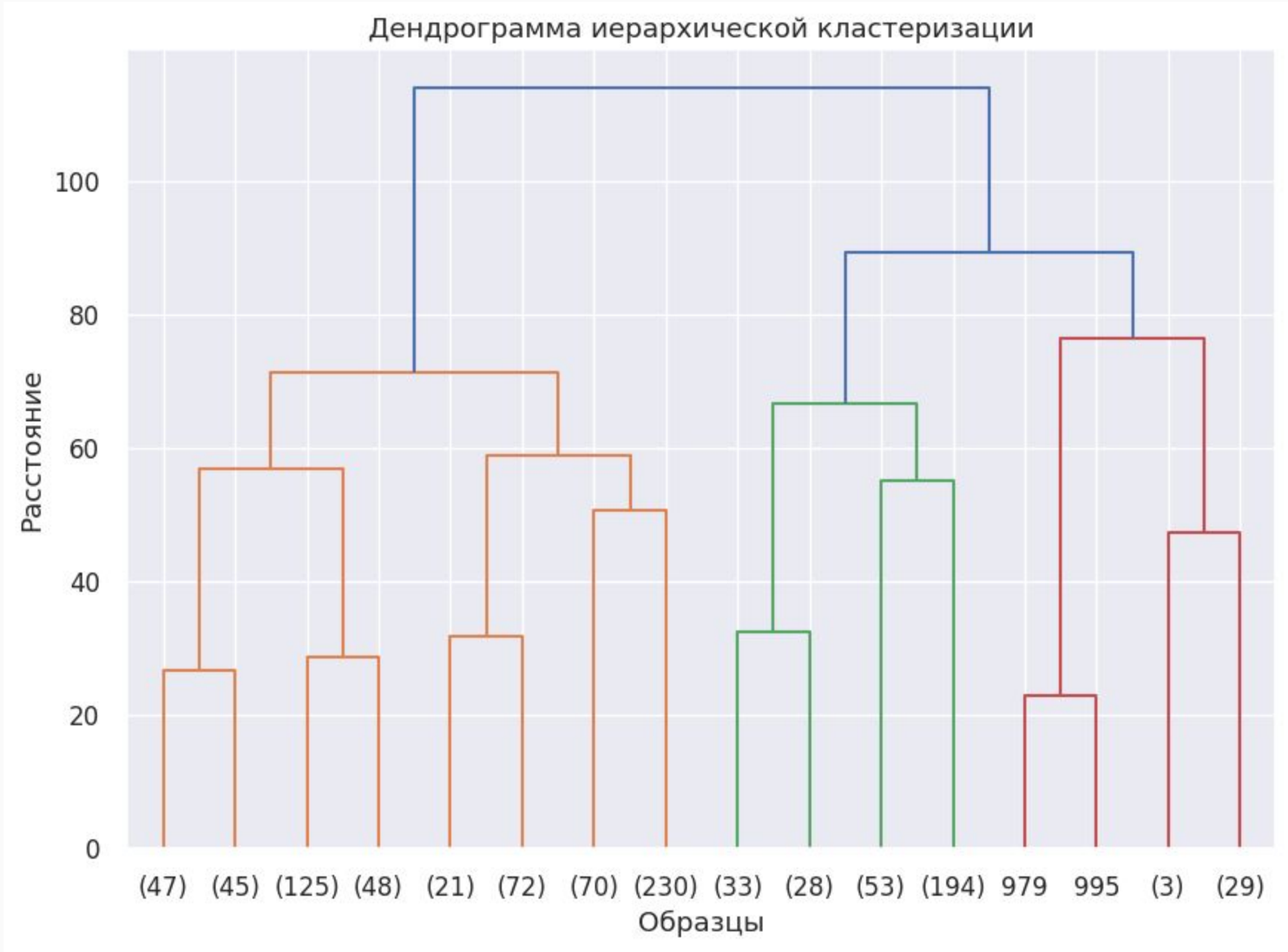
Сдвиги для SHAP значений



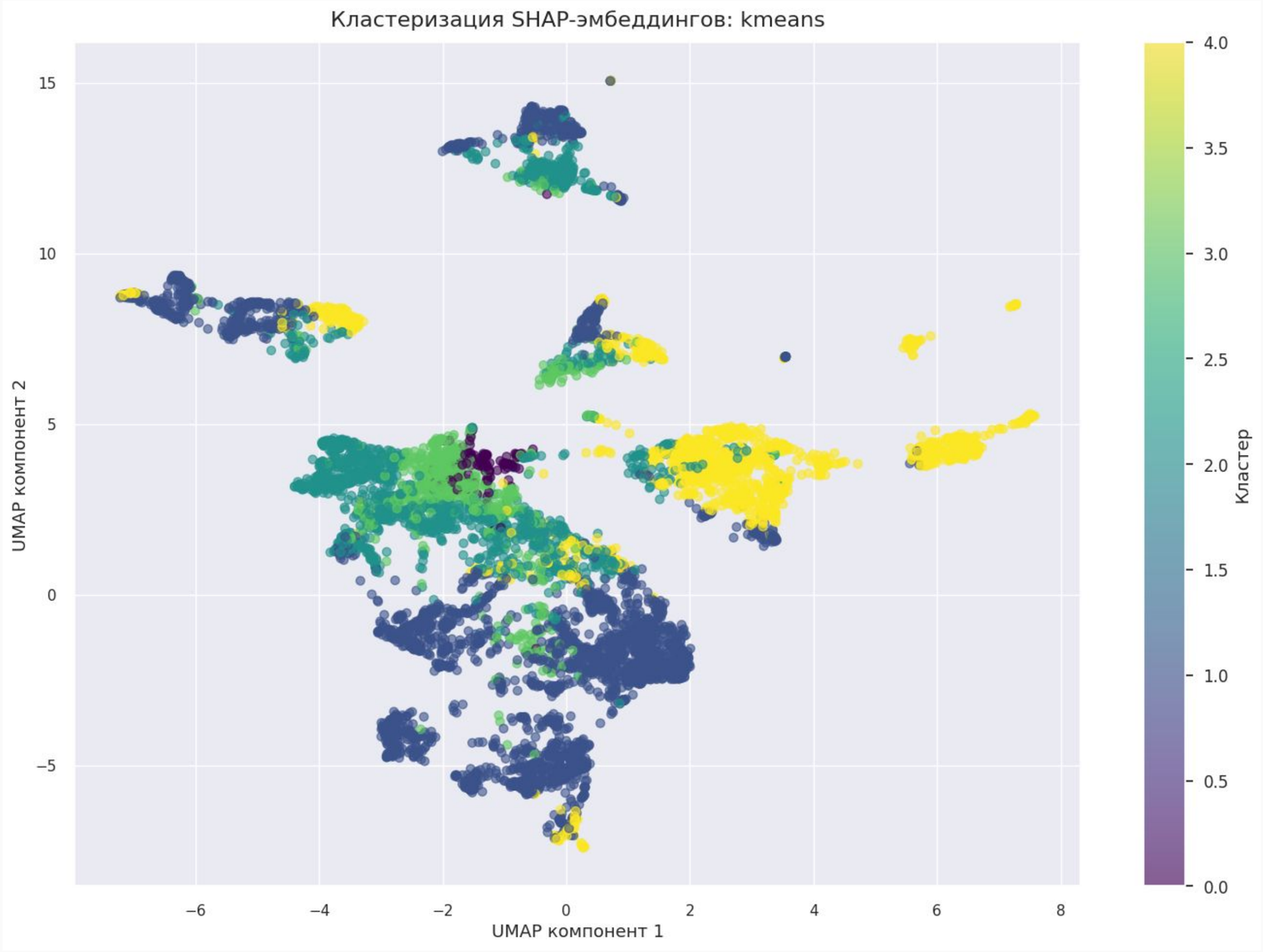
SHAP-аномалии



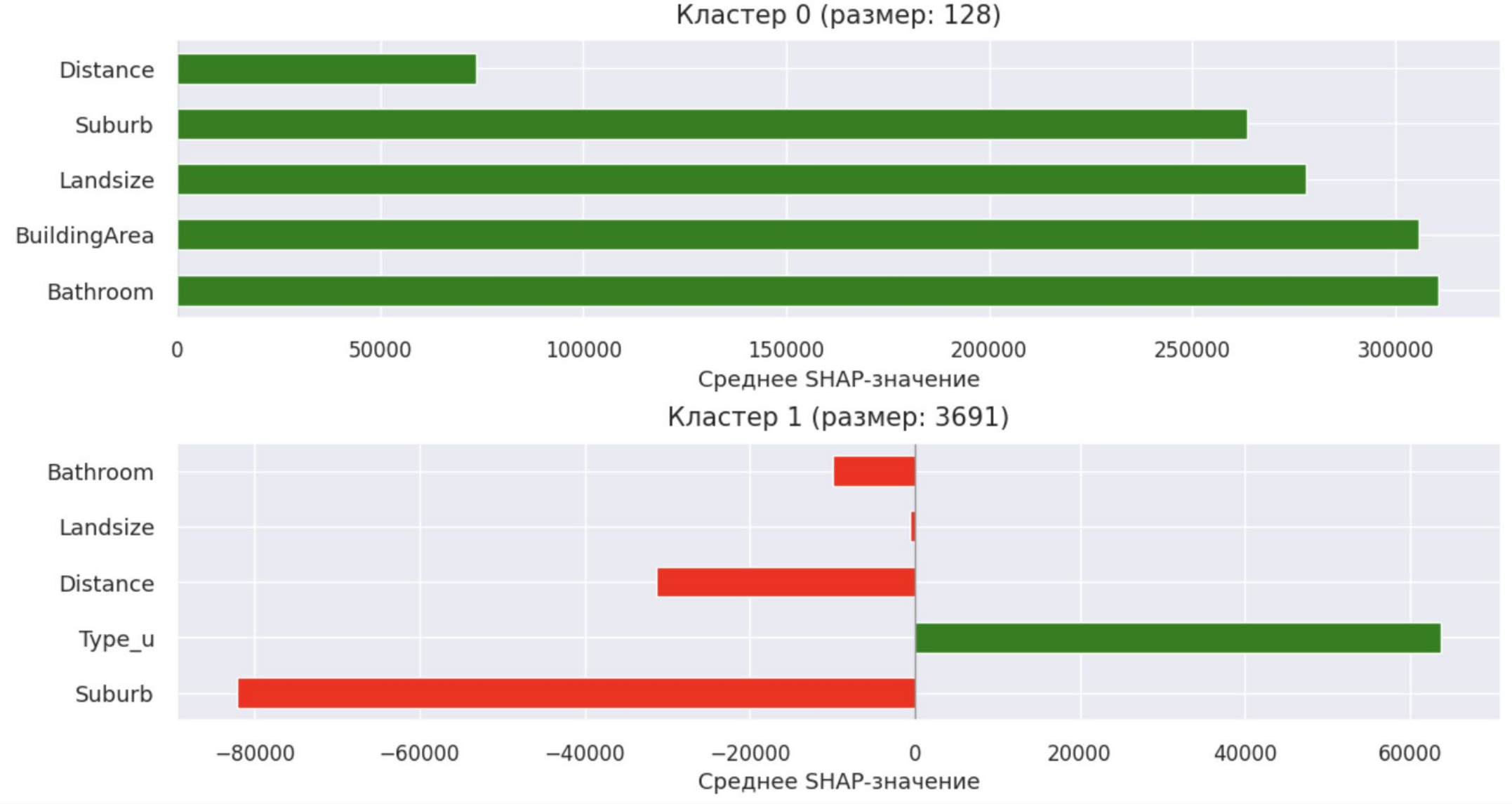
Дендрограмма и кластеры



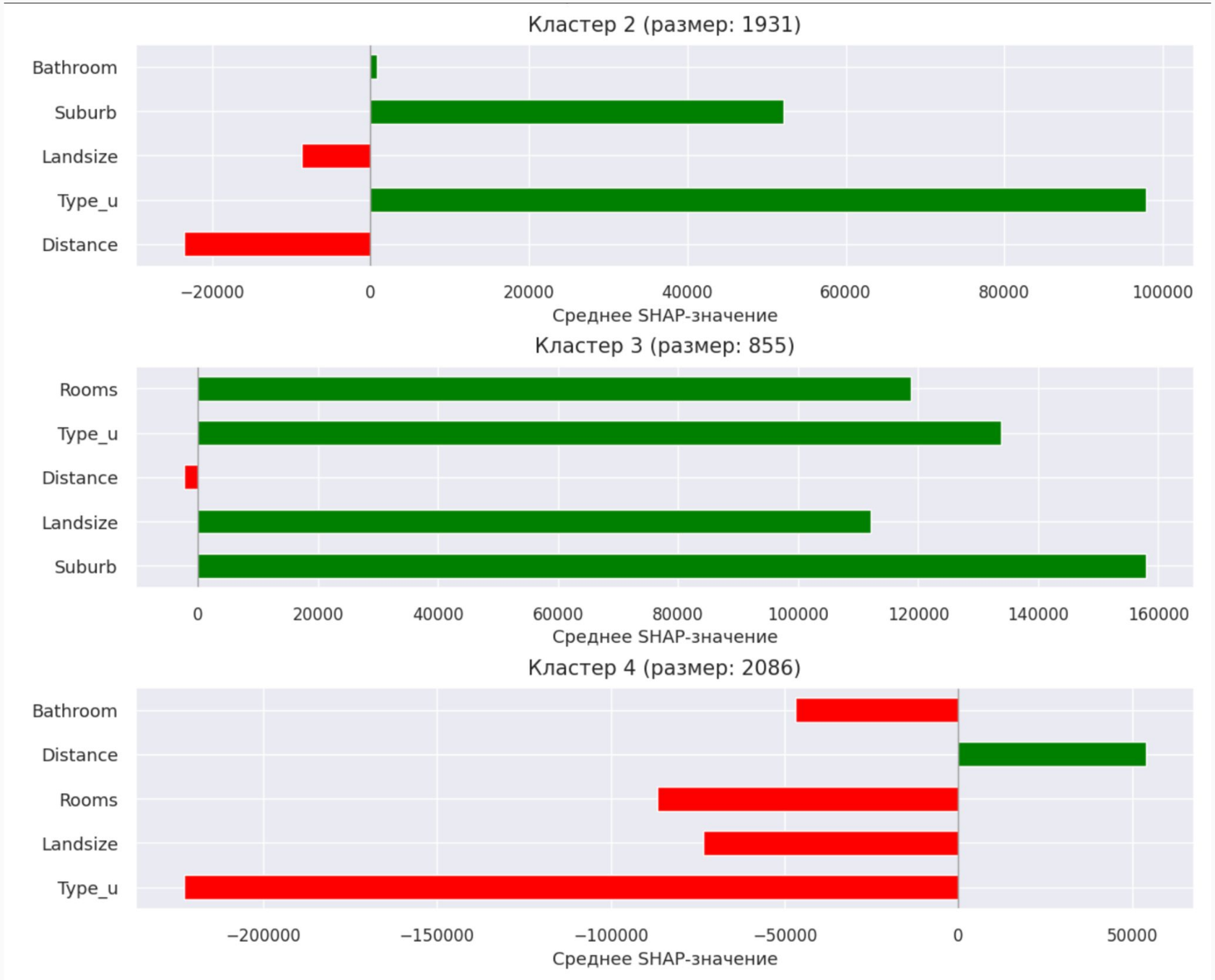
Кластеры Kmeans, Umap



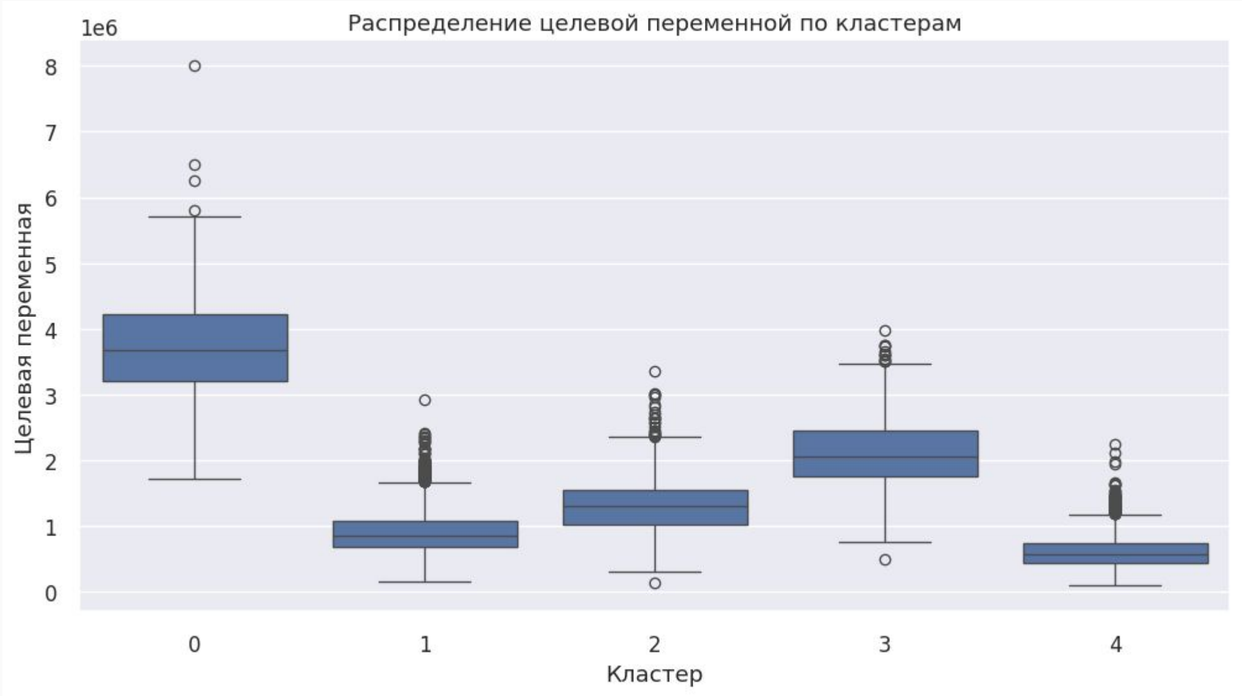
Кластеры

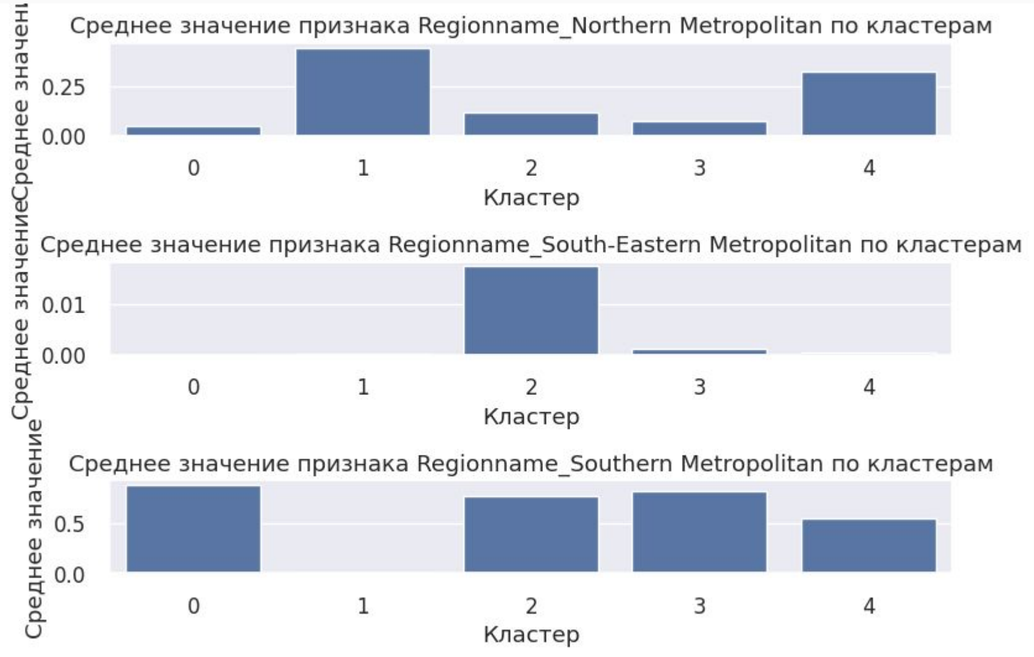

Кластер 0 (размер: 128)
Distance
Suburb
Landsize
BuildingArea
Bathroom
0
50000
100000
150000
200000
250000
300000
Среднее SHAP-значение
Кластер 1 (размер: 3691)
Bathroom
Landsize
Distance
Type_u
Suburb
−80000
−60000
−40000
−20000
0
20000
40000
60000
Среднее SHAP-значение

Кластеры

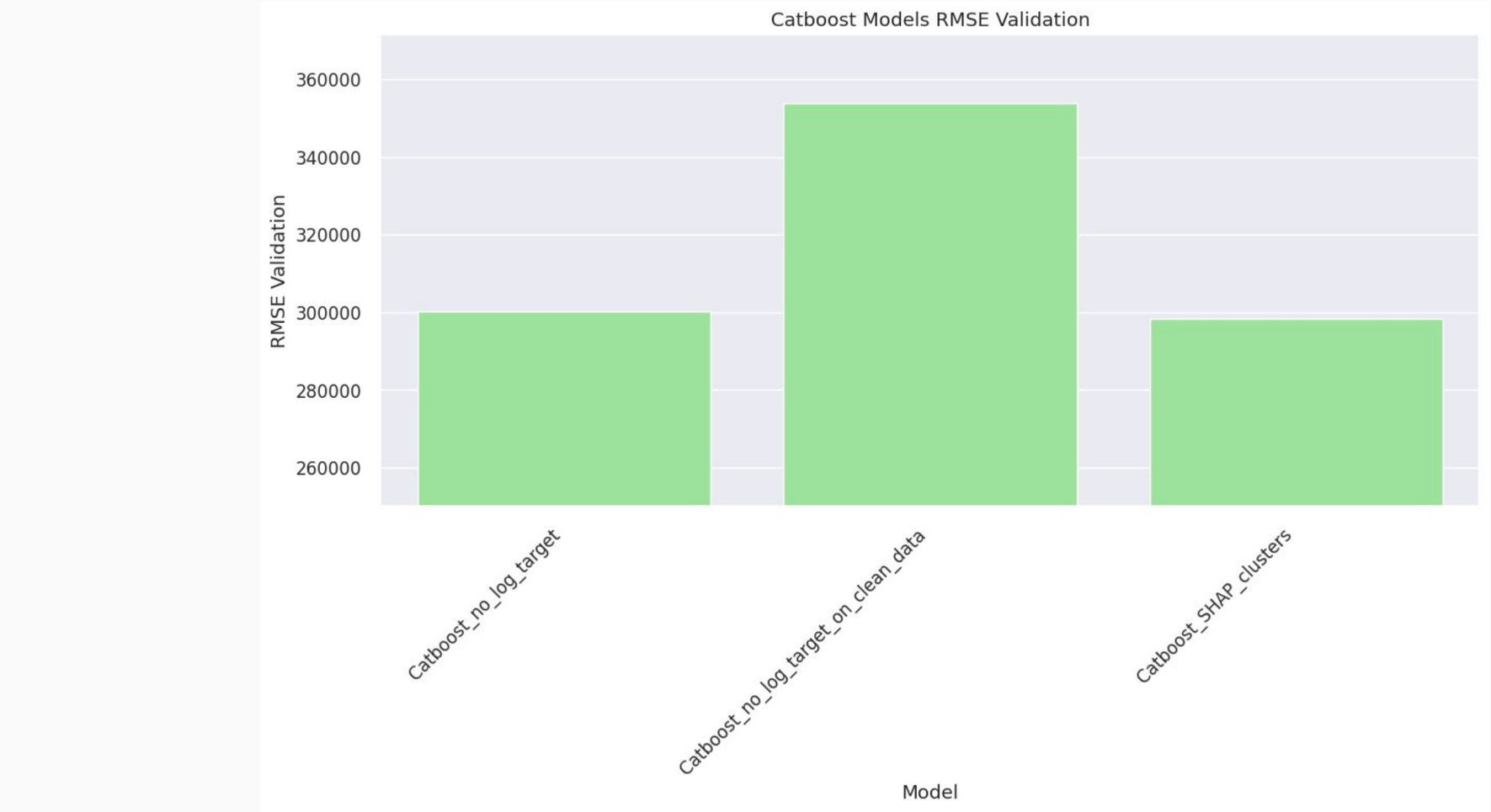

Кластер 2 (размер: 1931)
Bathroom
Suburb
Landsize
Type_u
Distance
−20000
0
20000
40000
60000
80000
100000
Среднее SHAP-значение
Кластер 3 (размер: 855)
Rooms
Type_u
Distance
Landsize
Suburb
0
20000
40000
60000
80000
100000
120000
140000
160000
Среднее SHAP-значение
Кластер 4 (размер: 2086)
Bathroom
Distance
Rooms
Landsize
Type_u
−200000
−150000
−100000
−50000
0
50000
Среднее SHAP-значение

Профили кластеров



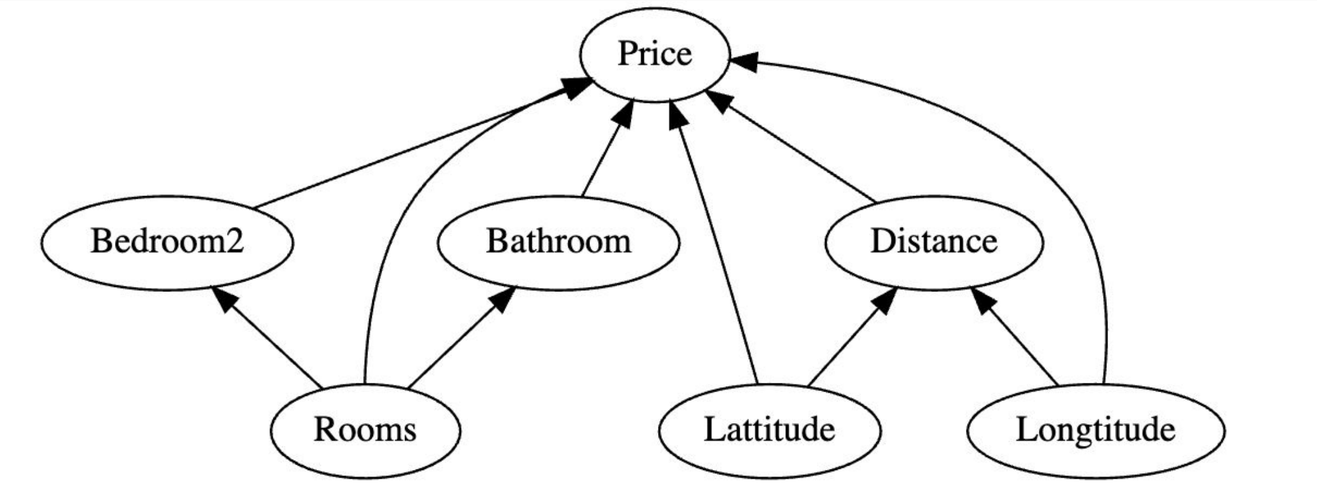


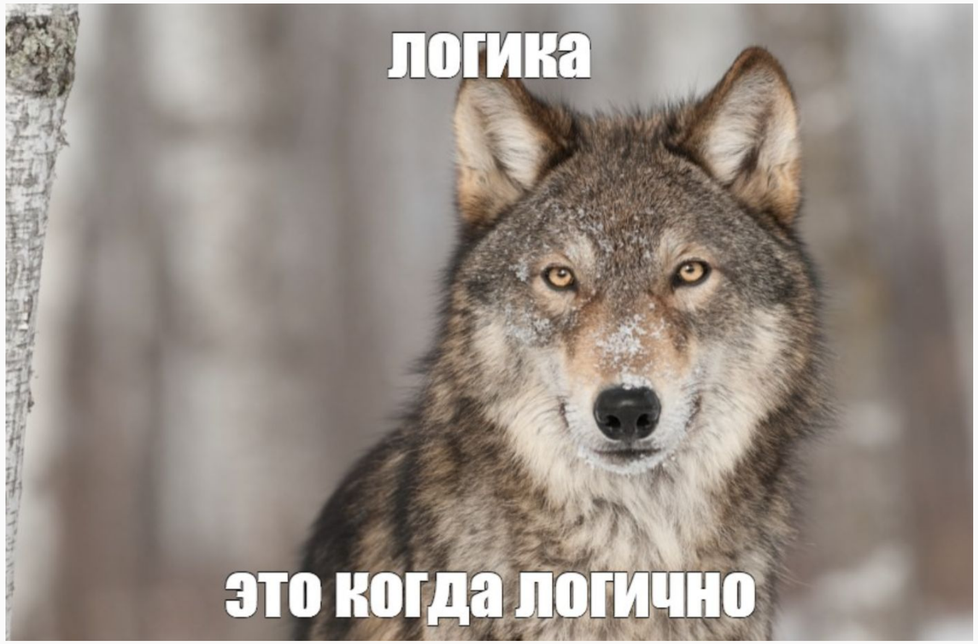
Анализ выбросов и кластеризация не помогают улучшить качество модели



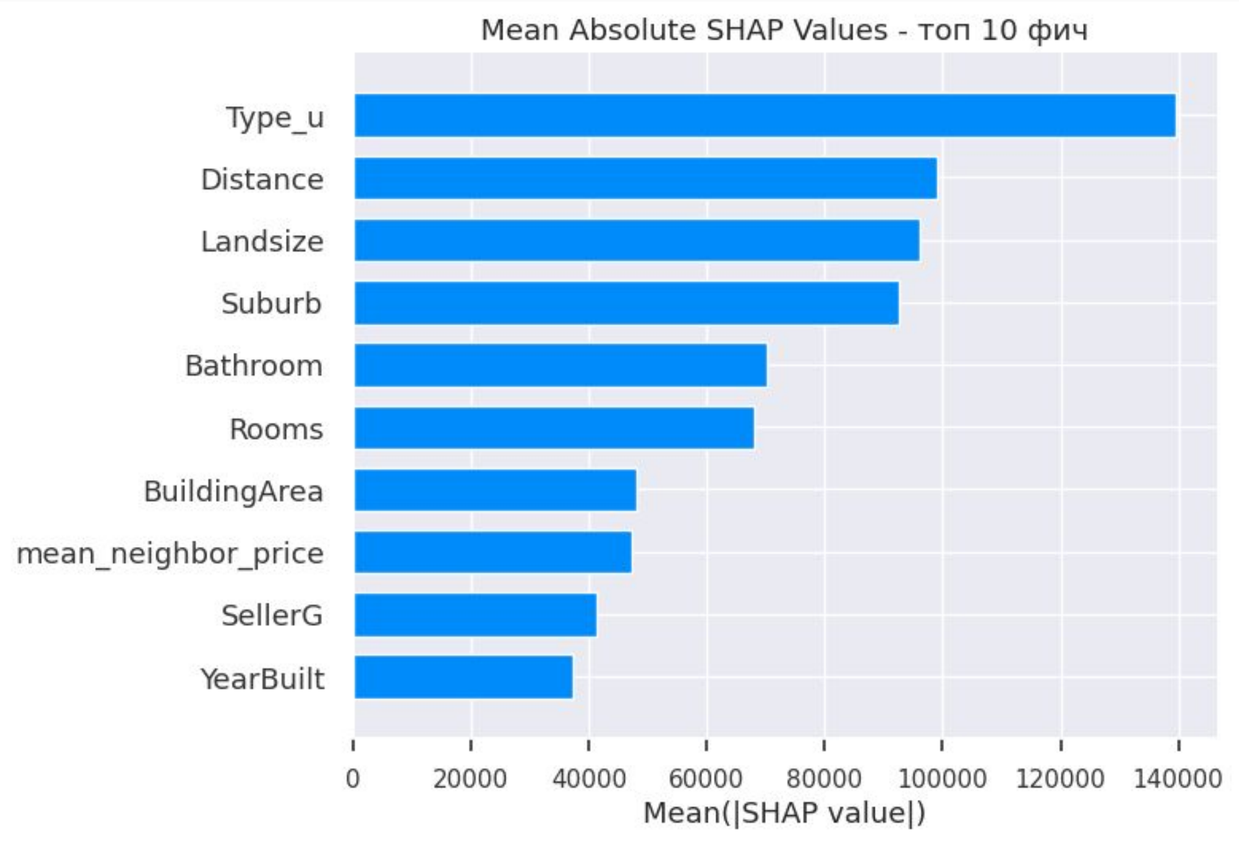
Shapley Flow

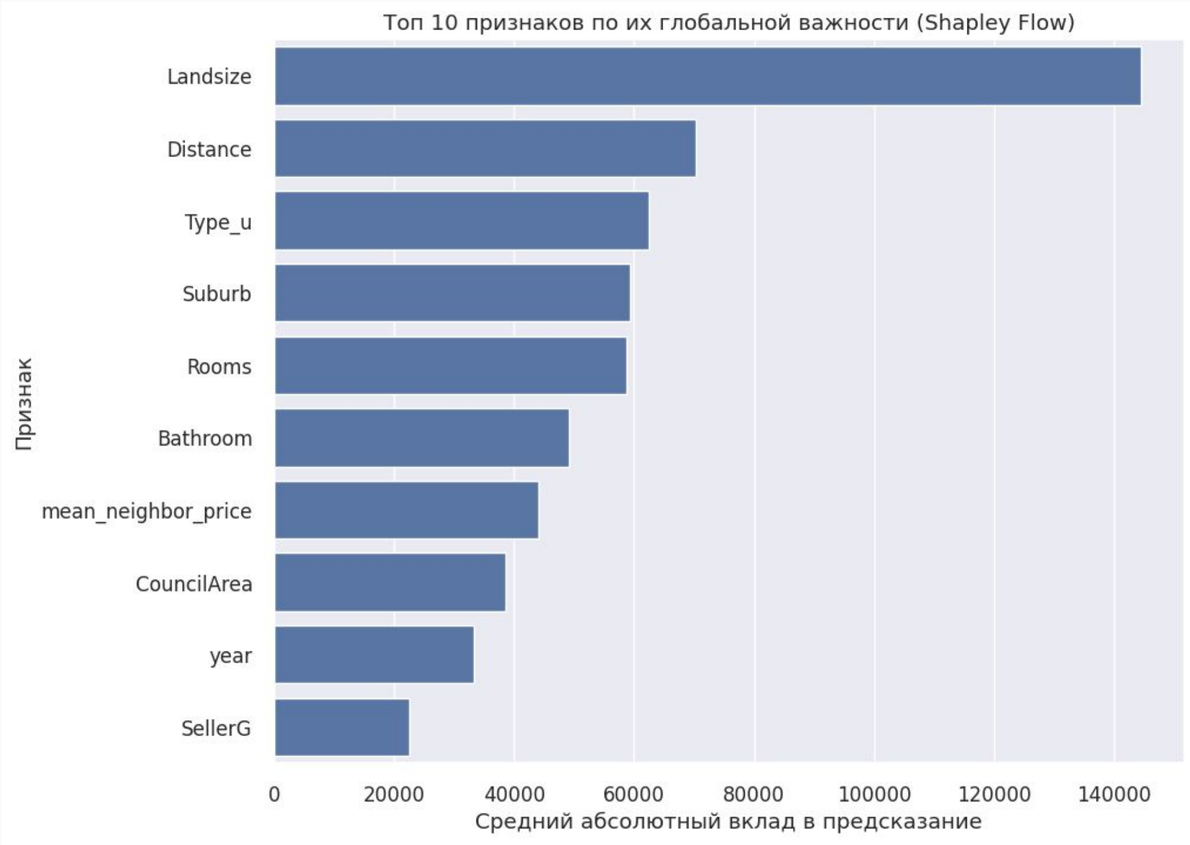
Для метода Shapley Flow нужно задать граф связей между переменными, мы используем самые очевидные связи



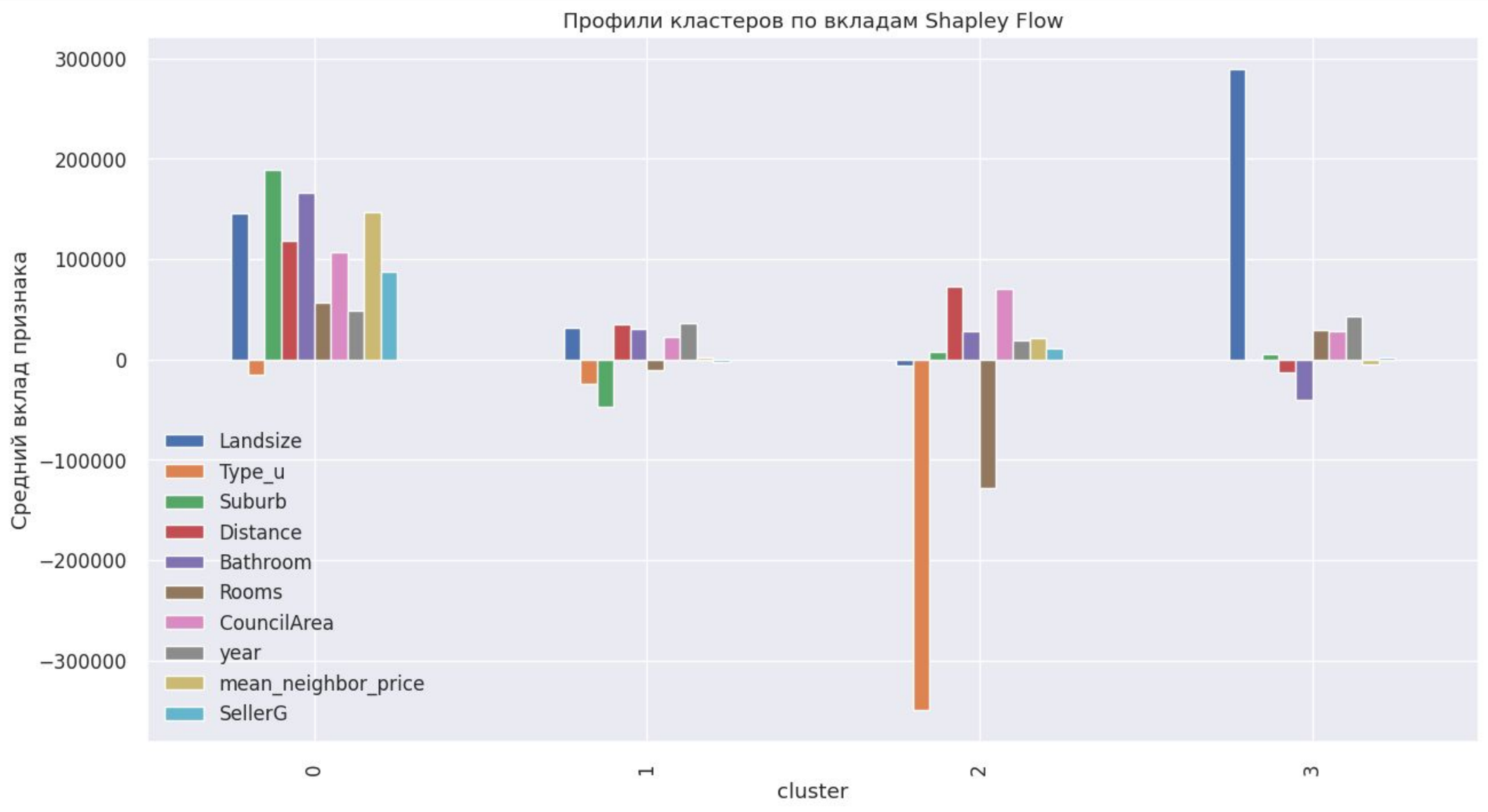


Сравнение самых важных признаков для SHAP и Shapley Flow

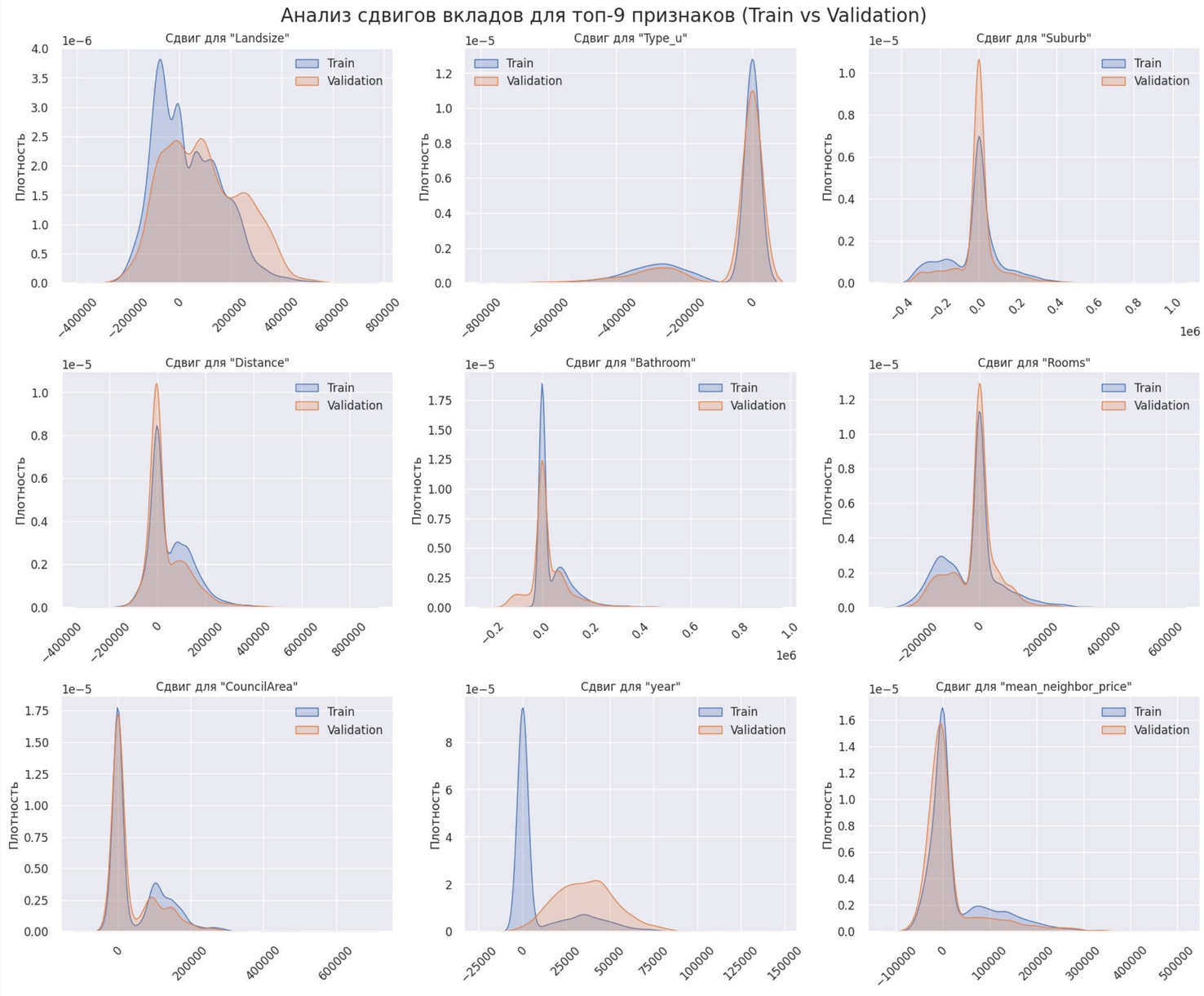




Кластеризация с помощью Shapely Flow



Сдвиги для Shapley Flow



Спасибо за внимание!