Отчет об анализе данных с сайта Циан.

1. Пропуски в данных

Количество записей в колонках сильно различается. Например, атрибуты, такие как станция метро, имеют много пропусков. Это нормально, учитывая что возле некоторых домов просто нет метро.

Ввод [2]:
```
missing_values = data.isnull().sum()
missing_values[missing_values > 0]
```

Out[2]:
```
agency_type              436
url                      436
region                   437
sale_type                440
property_type            440
floor                    442
total_floors             442
rooms                    442
area                     442
price_per_m2             443
is_mortgage_possible     445
total_price              447
district                3159
street                  3465
house_number            3828
metro_station           5201
residence_name          3735
dtype: int64
```

Ввод [3]:
```
numeric_columns = ['total_price', 'area', 'rooms', 'floor', 'total_floors']
for col in numeric_columns:
    data[col] = pd.to_numeric(data[col], errors='coerce')

data_cleaned = data.copy()
data_cleaned['price_per_m2'] = data_cleaned['total_price'] / data_cleaned['area']
```

Ввод [4]:
```
print(f"Количество записей: {len(data_cleaned)}")
```

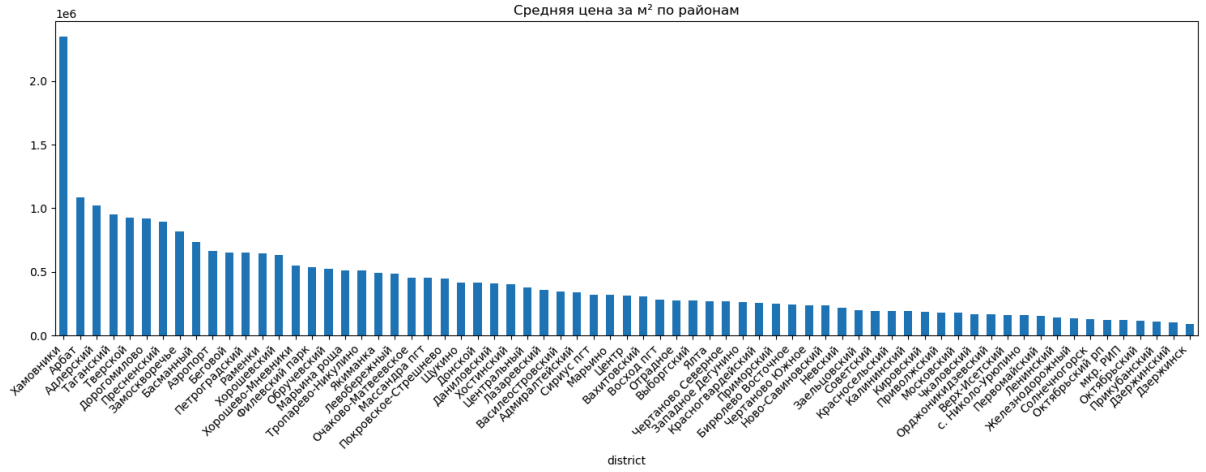
```
Количество записей: 11744
```

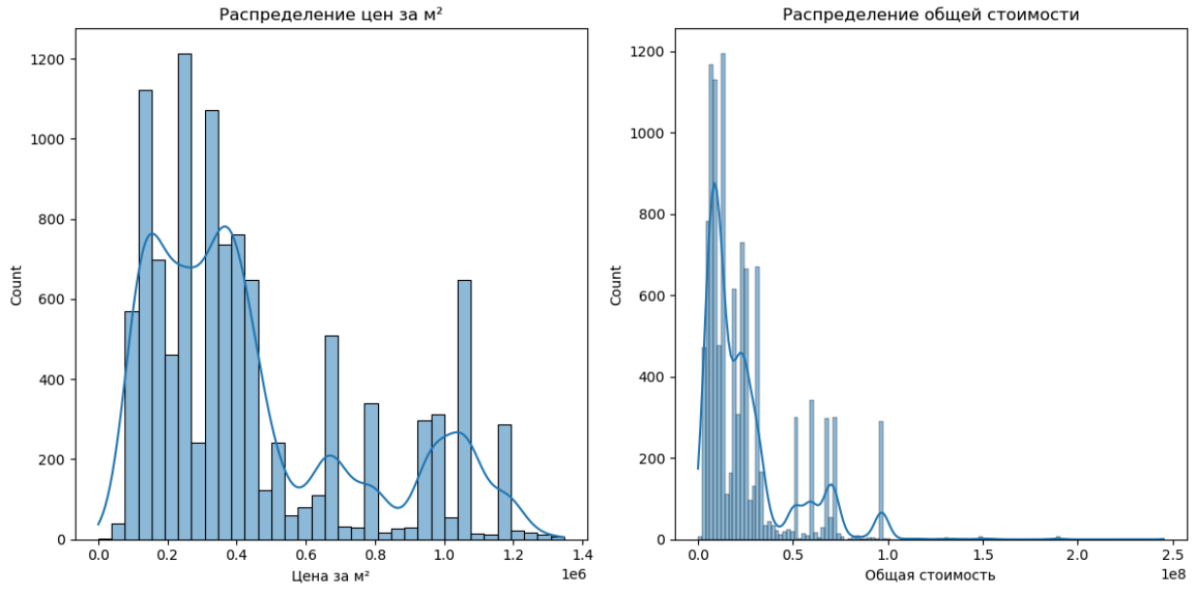
Ввод [5]:
```
data_cleaned.shape
```

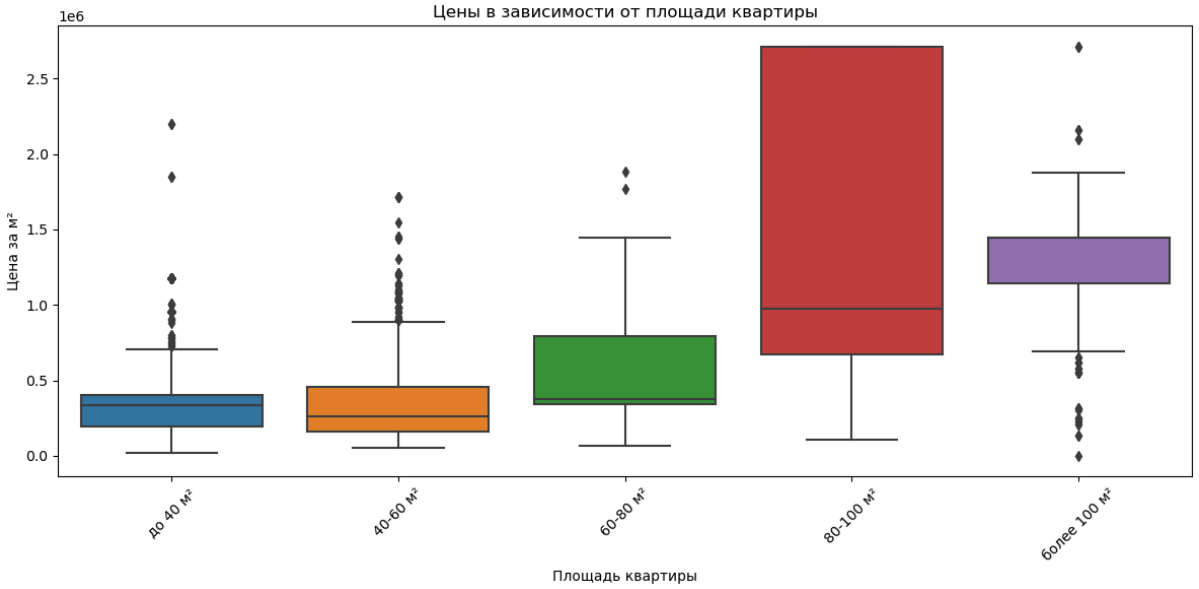
Out[5]:
```
(11744, 18)
```

2. Средняя цена за квадратный метр

- **По комнатности:** однокомнатные квартиры дешевле по цене за м², чем двухкомнатные. Это может объясняться высокой популярностью малогабаритных квартир, особенно среди покупателей с ограниченным бюджетом.
- **По этажности:** самые высокие цены за квадратный метр на средних этажах. Это логично, так как первый и последний этажи часто менее востребованы из-за шума, сырости, либо проблем с доступностью.
- **По площади:** цены за м² резко растут для квартир с площадью более 80 м². Возможно, это связано с тем, что большие квартиры находятся в более престижных районах или имеют лучшие характеристики, такие как вид из окон или дополнительные удобства.

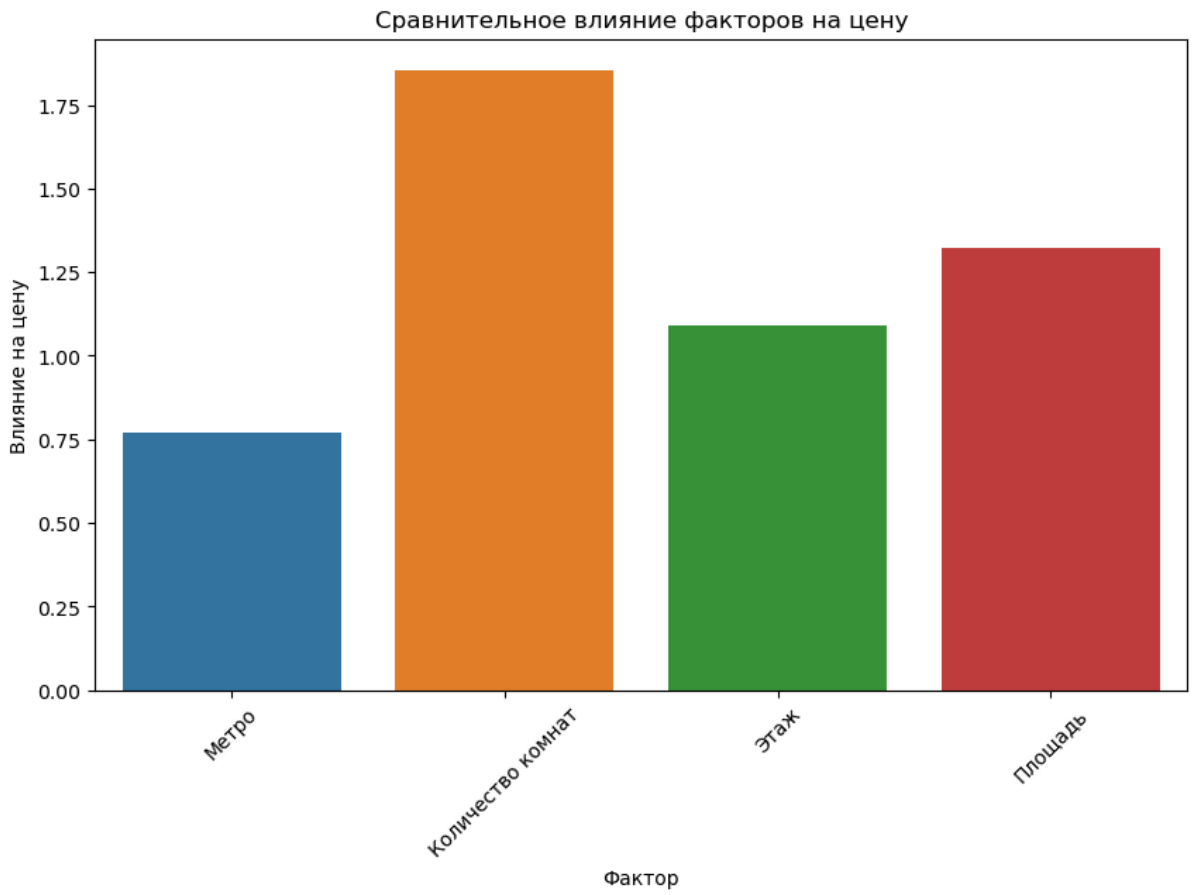






3. Влияние метро и этажности

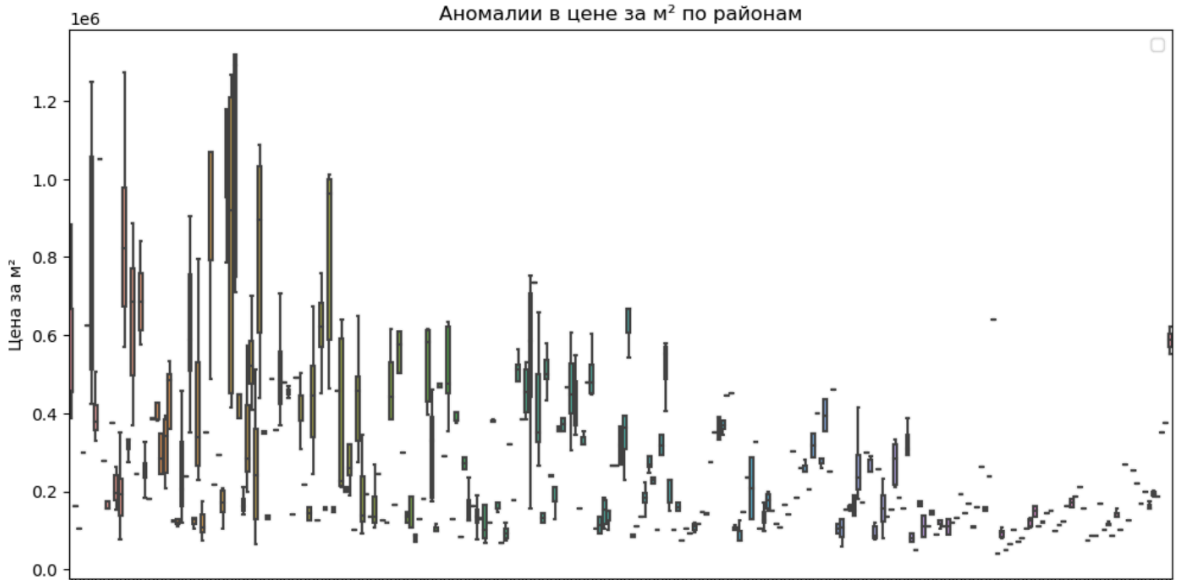
Присутствие рядом метро действительно сильно влияет на цену, увеличивая её почти на 77%. Это ожидаемо, так как наличие метро делает квартиру более привлекательной для широкого круга людей. Разница в цене между этажами также значительна — средние этажи оказываются предпочтительными.



ОСНОВНЫЕ ВЫВОДЫ:
1. КОРРЕЛЯЦИИ:
- total_price и rooms: 0.36
- total_price и price_per_m2: 0.82
- floor и total_floors: 0.65
2. НАИБОЛЕЕ ЗНАЧИМЫЕ ФАКТОРЫ:
- Количество комнат: 185.6% влияния на цену
- Площадь: 132.4% влияния на цену
- Этаж: 108.9% влияния на цену
- Метро: 76.9% влияния на цену
3. РАСПРЕДЕЛЕНИЕ ЦЕН:
- Медианная цена за м²: 346,617 руб
- 25% квартир дешевле: 198,376 руб/м²
- 75% квартир дешевле: 633,388 руб/м²

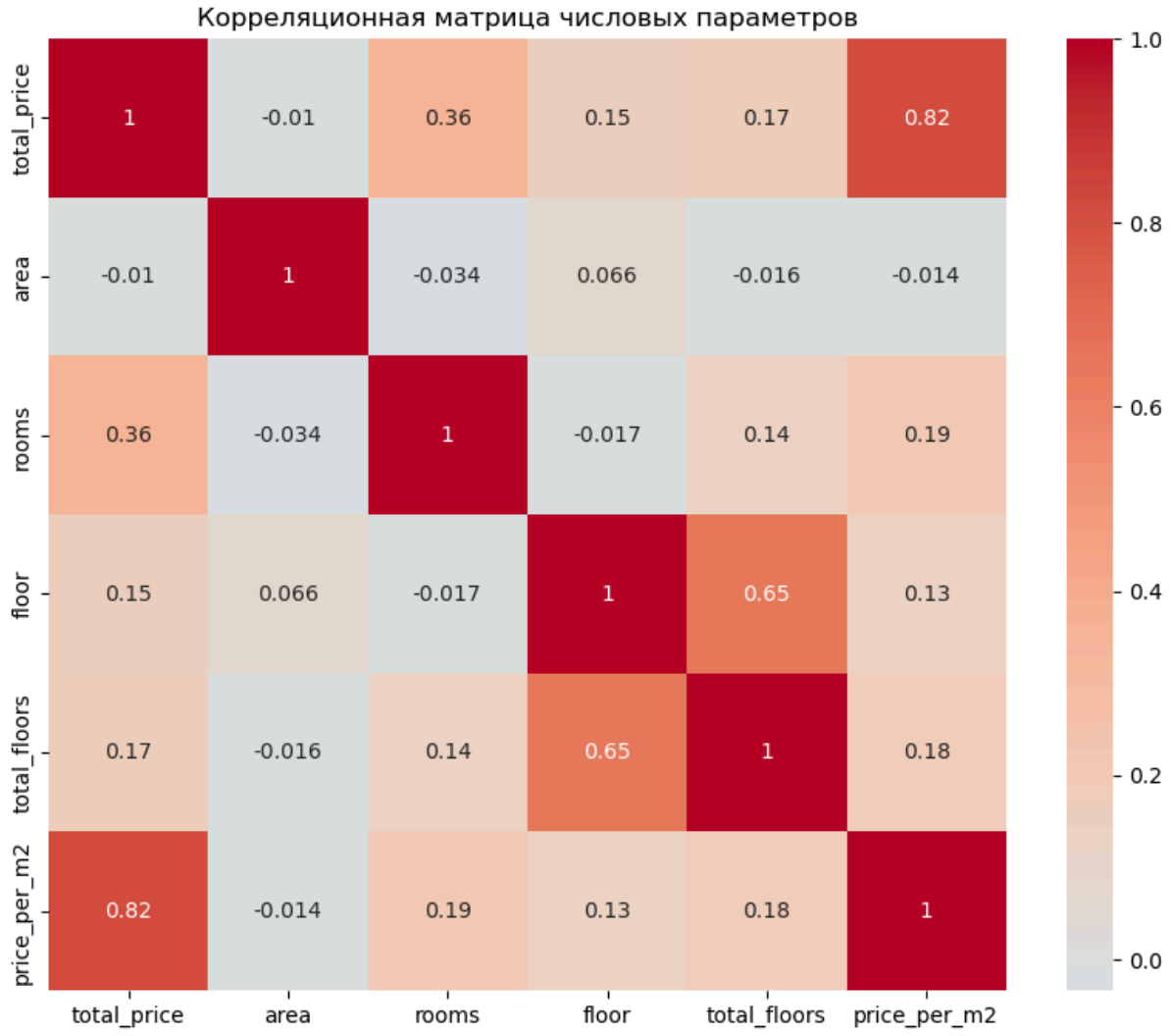
4. Аномалии в данных

Отсутствие аномалий по Z-оценке и наличие 894 аномалий по IQR говорит о том, что в данных есть выбросы, которые могут отражать как редкие, но реальные значения, так и ошибки. Это требует дополнительного анализа, чтобы решить, какие выбросы следует исключить.



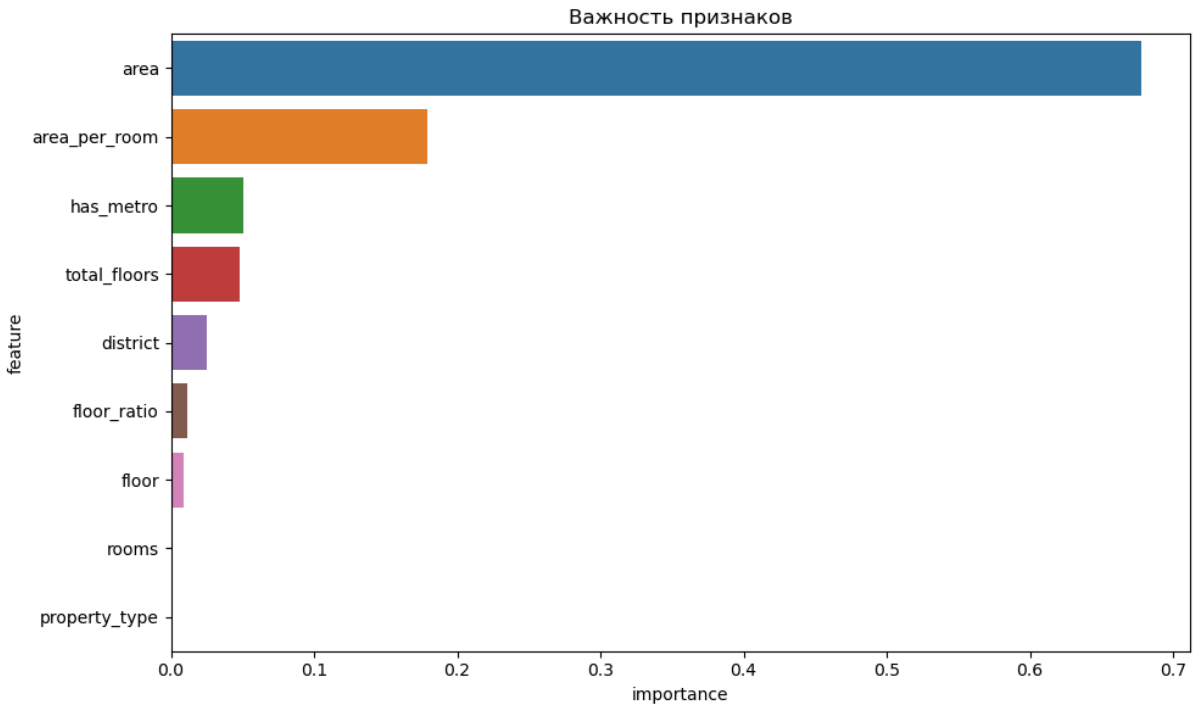
5. Корреляции

- Корреляция между ценой и количеством комнат невысока (0.36), что говорит о том, что цена зависит от других факторов, например, от площади.
- Высокая корреляция между ценой и ценой за м² (0.82) ожидаема, так как увеличение площади квартиры при фиксированной стоимости за м² увеличивает её общую стоимость.
- Корреляция между этажом и общим количеством этажей (0.65) может указывать на определенные особенности строения в регионе — например, более высокие дома могут быть более дорогими.



6. Влияние различных факторов на цену

Наибольшее влияние на цену оказывают количество комнат, площадь, этажность и близость к метро. Это подчеркивает значимость всех этих факторов при оценке стоимости квартиры.



7. Оценка моделей машинного обучения

- **Random Forest:** MAE и RMSE для модели достаточно хорошие, но заметно, что CatBoost работает значительно точнее.
- **CatBoost:** значительно более низкие значения MAE и RMSE и высокий R^2 (0.992) говорят о том, что модель CatBoost лучше подходит для данной задачи. Ошибки на уровне 7-10% говорят о высокой точности модели.

Обучение модели Random Forest...

Метрики Random Forest:

MAE: 29,977 руб/м²

RMSE: 88,317 руб/м²

R2 Score: 0.966

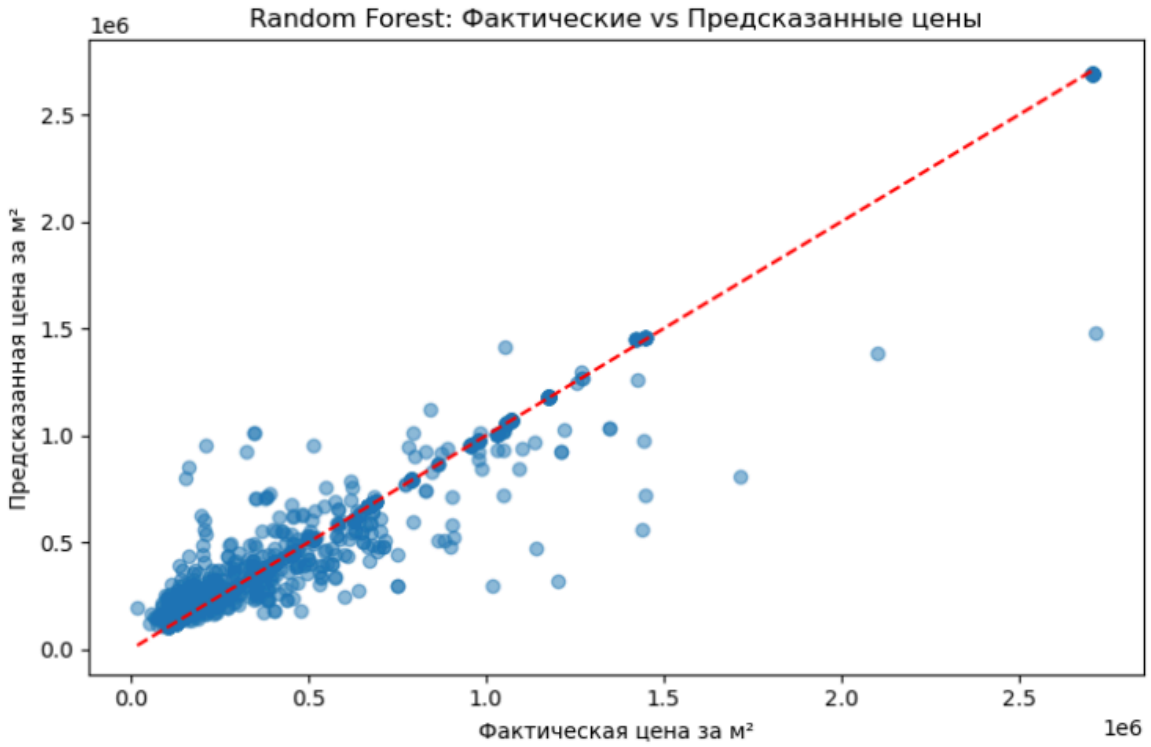
Обучение модели CatBoost...

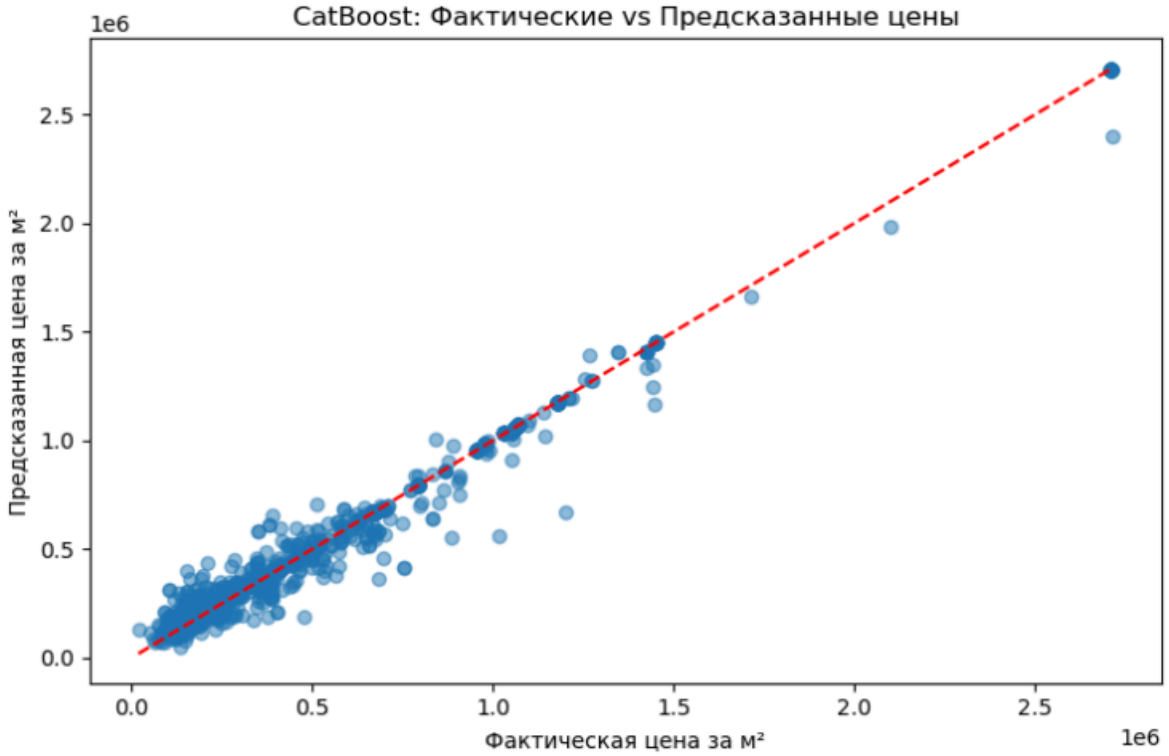
Метрики CatBoost:

MAE: 16,211 руб/м²

RMSE: 43,335 руб/м²

R2 Score: 0.992





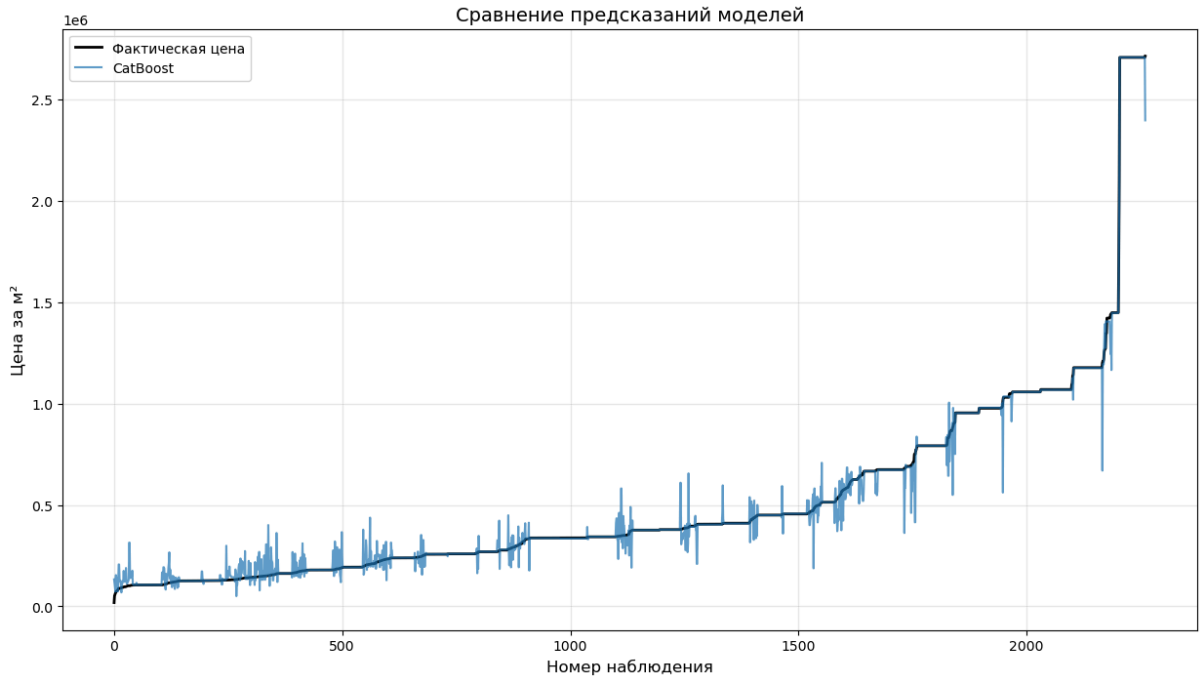
8. Примеры предсказаний (CatBoost)

Ошибки в предсказаниях по конкретным примерам незначительны, особенно для дорогих квартир. Это говорит о том, что модель хорошо справляется с предсказаниями на больших значениях.

Ввод [24]:

```python
# Примеры конкретных предсказаний
print("Примеры предсказаний (CatBoost):")
sample_indices = np.random.choice(len(y_test), 5)
for idx in sample_indices:
    actual = y_test.iloc[idx]
    predicted = predictions['CatBoost'][idx]
    print(f"Фактическая цена: {actual:,.0f} руб/м²")
    print(f"Предсказанная цена: {predicted:,.0f} руб/м²")
    print(f"Разница: {abs(actual - predicted):,.0f} руб/м²")
    print(f"Относительная ошибка: {abs(actual - predicted)/actual*100:.1f}%")
```

```
Примеры предсказаний (CatBoost):
Фактическая цена: 602,326 руб/м²
Предсказанная цена: 557,167 руб/м²
Разница: 45,158 руб/м²
Относительная ошибка: 7.5%
Фактическая цена: 1,177,500 руб/м²
Предсказанная цена: 1,175,841 руб/м²
Разница: 1,659 руб/м²
Относительная ошибка: 0.1%
Фактическая цена: 145,000 руб/м²
Предсказанная цена: 160,654 руб/м²
Разница: 15,654 руб/м²
Относительная ошибка: 10.8%
Фактическая цена: 1,069,678 руб/м²
Предсказанная цена: 1,067,766 руб/м²
Разница: 1,912 руб/м²
Относительная ошибка: 0.2%
Фактическая цена: 108,893 руб/м²
Предсказанная цена: 100,204 руб/м²
Разница: 8,689 руб/м²
Относительная ошибка: 8.0%
```



Общий вывод

Анализ данных показывает, что на цену недвижимости в значительной степени влияет этаж, близость к метро, общая площадь и количество комнат. Модели машинного обучения, особенно CatBoost, показали себя очень эффективными в предсказании стоимости.