# ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ОЦЕНКИ СРОКА ЭКСПОЗИЦИИ ОБЪЕКТОВ КОММЕРЧЕСКОЙ НЕДВИЖИМОСТИ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Цель данного исследования: оценить возможность определения сроков экспозиции отдельных объектов, используя в качестве предикторов количество просмотров в единицу времени. Это позволит более точно предсказывать ожидаемый срок экспозиции по сегментам недвижимости с привязкой к конкретным зонам нахождения объектов.

## Предобработка

Для моделирования на первом этапе были использованы данные move по офисной недвижимости. Далее было принято решение использовать данные по всем сегментам коммерческой недвижимости также из move.
На этапе предобработки были выполнены следующие шаги:

1. Удаление строк с пропущенными значениями.
2. Преобразование признака Общая площадь в категориальный Группа площади. Итоговые группы: <50, 50-100, 100-500, 500-1000, 1000-3000, >3000 м$^2$.
3. Группировка городов миллионников в единую группу по признаку Город. Новая группа так и называется - Миллионники.
4. Удаление объявлений, которые не были сняты с продажи из выборки.
5. Создание новых признаков на основе имеющихся:
    a. Был создан признак Кол-во просмотров в ед. времени = Кол-во просмотров / Срок экспозиции.
    b. Был создан признак Интервальное кол-во просмотров в ед. времени, изначальный числовой признак был преобразован в категориальный с интервалами от 0 до 1 с шагом в 0.2. Все значения больше 1 были объединены в одну группу (0-0.2, 0.2-0.4, …, 0.8-1, >1).
    c. Целевой признак Срок экспозиции был преобразован в категориальный с шагом в 30 дней (группы 0-29, 30-59, … 390-419, >420) и назван Интервальный Срок экспозиции.
    d. Значения Срока экспозиции были округлены до ближайшего числа кратного 30 в большую сторону (0-30 -> 30, 31-60 -> 60 и т.д.), новый признак - Округленный Срок экспозиции.

## Подготовка данных к обучению

Как было сказано ранее изначально для моделирования использовались данные только по офисной недвижимости в кол-ве 77041 наблюдения. Далее была использована выборка, в которой 343786 наблюдений.

Для моделирования данные были разбиты на тренировочную, валидационную и тестовые выборки в соотношении 0.7/0.15/0.15.

## Моделирование

На этапе моделирования были построены несколько моделей с использованием алгоритма Градиентного бустинга библиотеки CatBoost.

Изначально была построена модель, использовавшая данные по офисной недвижимости и предикторы: Город, Класс, Удельная цена, Ценовая зона, Группа площади, Кол-во просмотров. Результат по целевой метрике MdAPE(Медиана абсолютной процентной ошибки) составил 18.92 на тренировочной выборке и 19.56 на тестовой выборке и по диаграмме рассеяния можно сказать, что в целом модель имеет хорошую прогнозную способность:

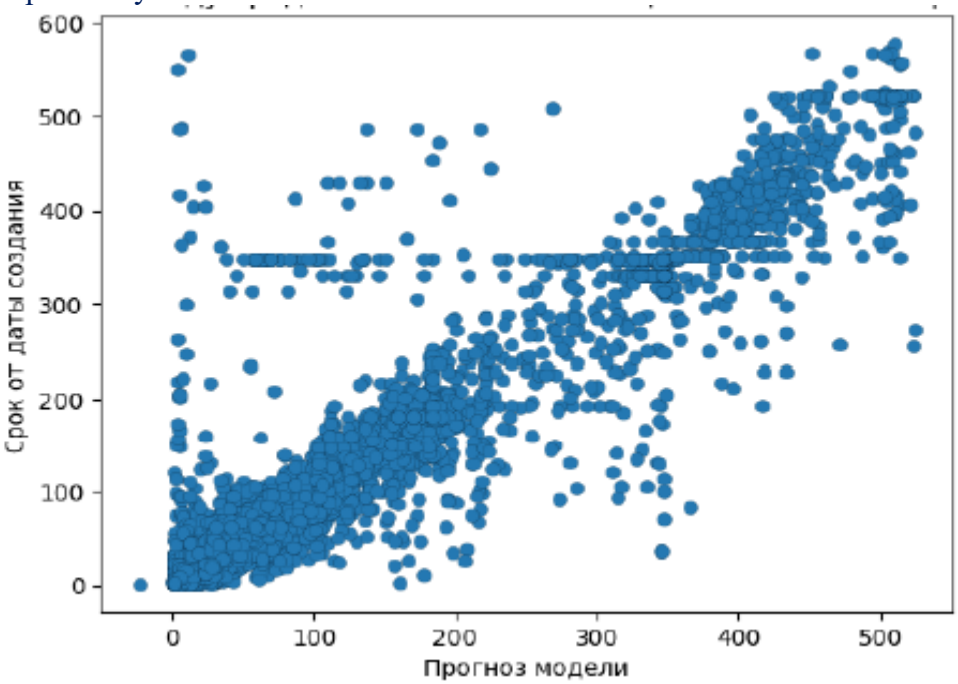


*Рисунок 1. Диаграмма рассеяния между предсказанными и настоящими значениями срока экспозиции, модель 1*

Однако, такой способ моделирования не учитывает темпы получения объявлениями просмотров, поскольку общее кол-во просмотров для одного объявления может быть, например, 100 за 1 день, а для другого 100 за 10 дней и это будет иметь одинаковый вес. Было принято решение использовать в качестве предиктора признак Кол-во просмотров в ед. времени вместо Кол-ва просмотров. Для построения новой модели был вычислен показатель кол-ва просмотров в день для тренировочной выборки и обучена модель, с помощью которой данный показатель был спрогнозирован для тестовой выборки. MdAPE для модели, прогнозирующей Кол-во просмотров в ед. времени, составил 24.51. Также, для тренировочных данных была оценена зависимость срока экспозиции от кол-ва просмотров в ед. времени и можно сделать вывод, что линейной зависимости не наблюдается:

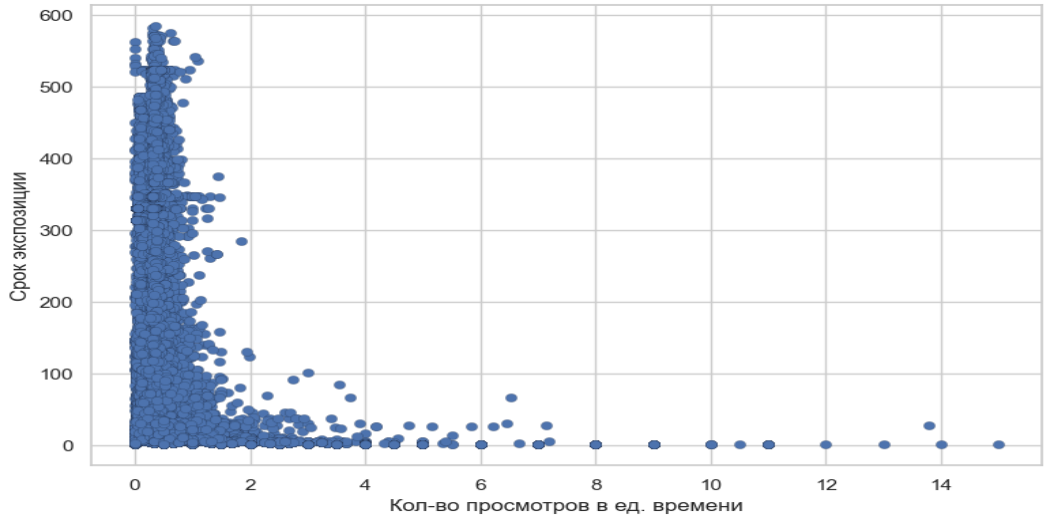


*Рисунок 2. Зависимость срока экспозиции от кол-ва просмотров, тренировочная выборка, настоящие значения*

Таким образом для обучения второй модели прогнозирующей срок экспозиции были использованы те же самые предикторы, что и для первой, кроме Кол-ва просмотров, который был заменен на признак Кол-во просмотров в ед. времени. Итоговое значение метрики MdAPE составило 49,44 на тренировочной выборке и 73,75 на тестовой. Диаграмма рассеяния также подтверждает ухудшение прогнозной способности модели:

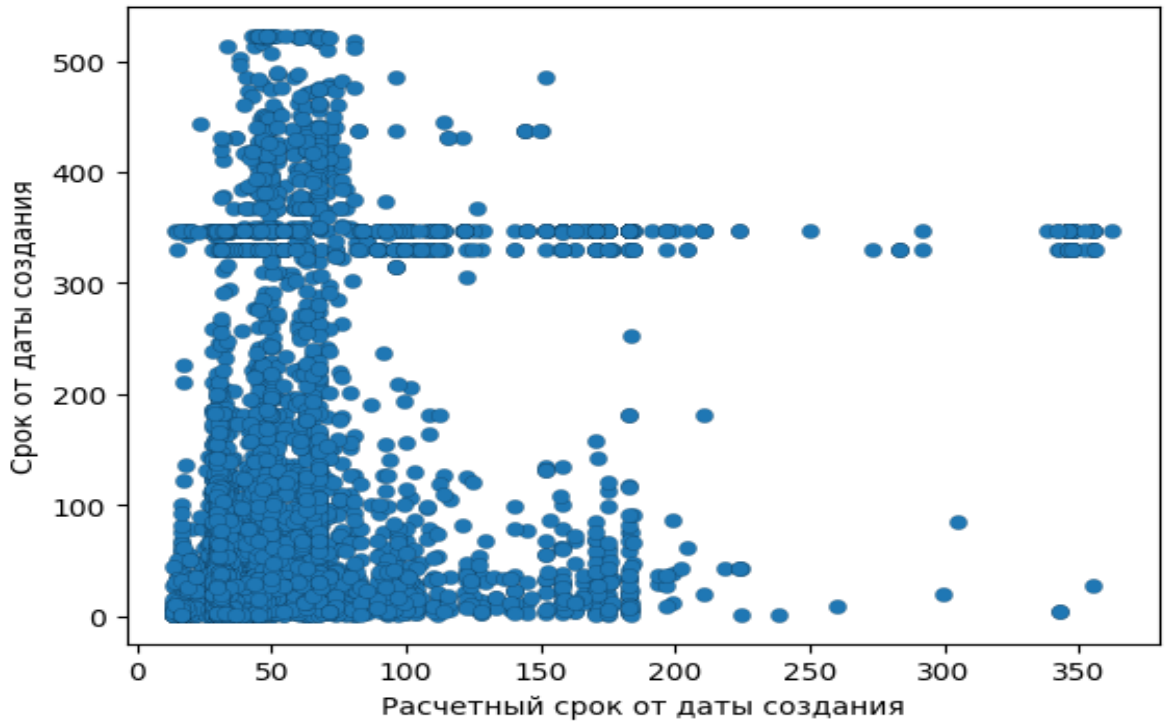


*Рисунок 3. Диаграмма рассеяния между предсказанными и настоящими значениями срока экспозиции, модель 2*

Так как результаты после использования нового признака стали заметно хуже было принято решение попробовать использовать интервальное значение кол-ва просмотров, в качестве предиктора. Кроме того, было принято решение использовать

для обучения и тестирования моделей не только объявления по офисной недвижимости и но и другие имеющиеся объявления из коммерческого сегмента. Срок экспозиции был вновь оценен и значение метрики MdAPE немного улучшилось, став равным 53.83 на тренировочной и 66.38 на тестовой выборках:

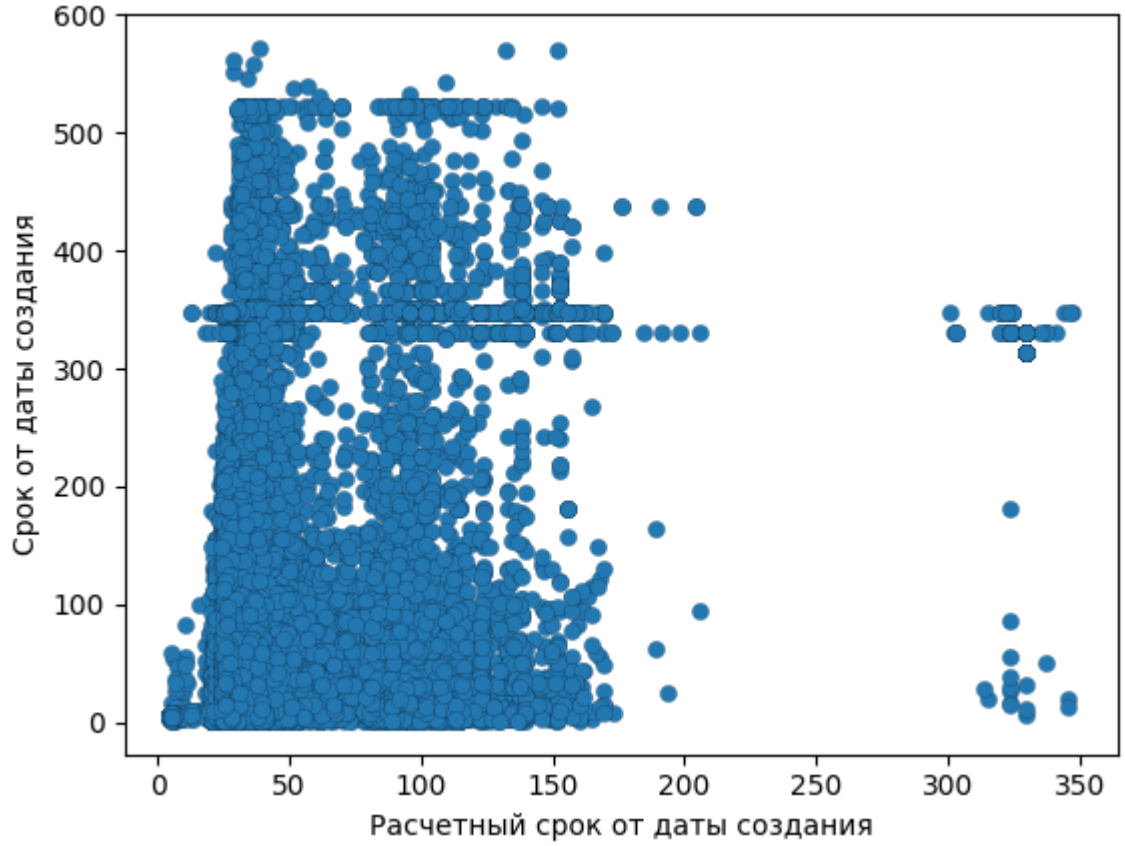


*Рисунок 4. Диаграмма рассеяния между предсказанными и настоящими значениями срока экспозиции, модель 3*

Поскольку фактический срок экспозиции имеет чрезвычайно большой разброс, при проведении дальнейших исследований использовались медианные значения сроков экспозиции (фактические и предсказанные). С этой целью вся область изменения значений количества просмотров в единицу времени разделена на интервалы: (0-0.2), (0.2-0.4), …, (0.8-1), (1- >1). Всего 6 интервалов. Для каждого интервала определялось медианное значение количества просмотров в единицу времени. Соответственно для значений сроков экспозиции по объектам, относящимся к этим интервалам, рассчитывались медианные сроки экспозиции. Таким образом, были получены 6 пар значений для количества просмотров в единицу времени и сроков экспозиции. Результаты представлены на графиках (рис.5 и рис.6)

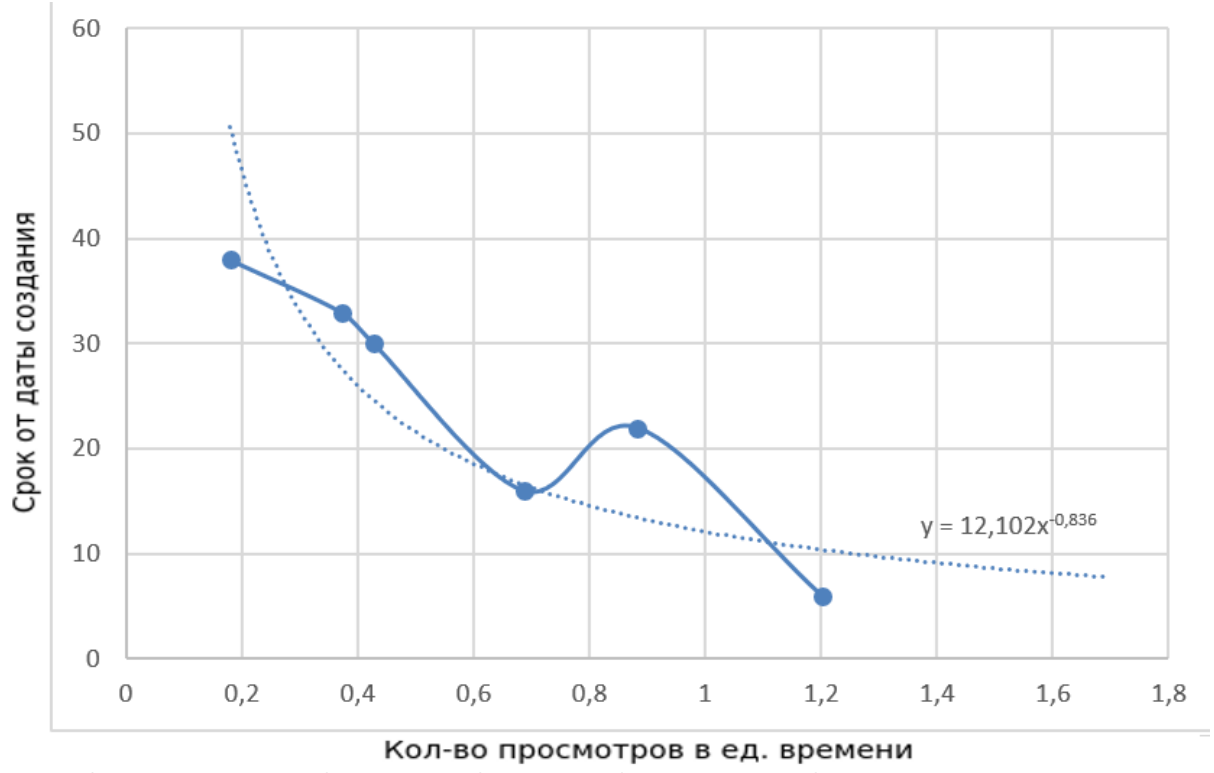


*Рисунок 5. Зависимость медианы срока экспозиции от медианы кол-ва просмотров внутри групп, тестовая выборка, настоящие значения срока*

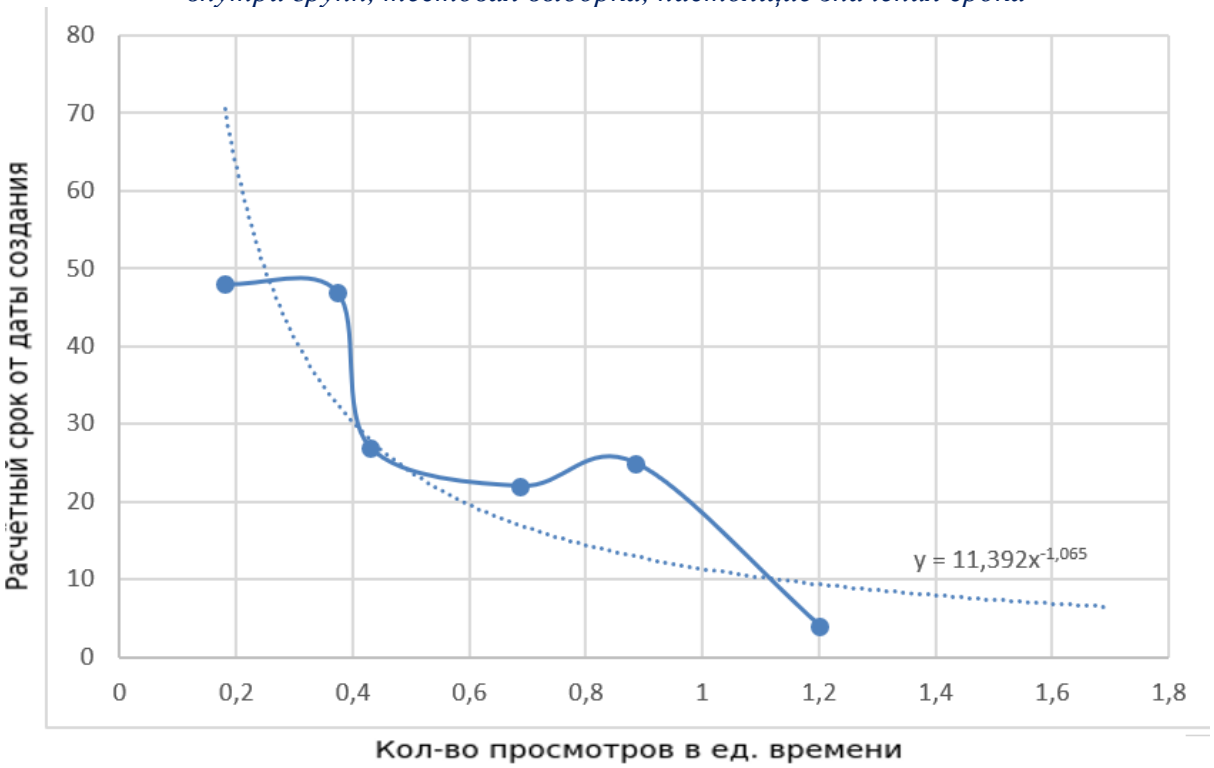


*Рисунок 6. Зависимость медианы срока экспозиции от медианы кол-ва просмотров внутри групп, тестовая выборка, прогнозные значения срока*

Как можно видеть из графиков, имеется четко выраженная зависимость, между сроком экспозиции и кол-вом просмотров. Причем эта зависимость сохраняется как для настоящих значений срока, так и для расчетных. Взаимосвязь также подтверждается графиком важности признаков CatBoost:

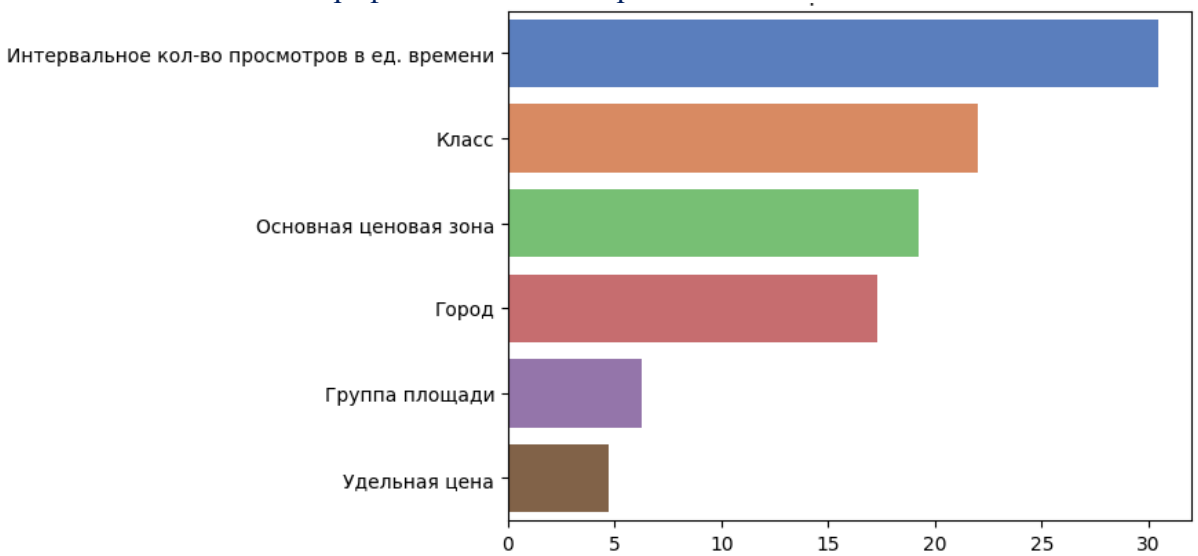


*Рисунок 7. Важность признаков CatBoost, модель 3*

Приведенные графики показывают, что Интервальное кол-во просмотров является наиболее важным признаком из представленных.

На финальном этапе моделирования, для улучшения результата, было принято решение спрогнозировать не сам срок экспозиции, а его категориальный вариант Интервальный Срок экспозиции и числовой Округленный Срок экспозиции с округлением до ближайшего числа кратного 30 в большую сторону. В первом случае задача изменилась с регрессии на многоклассовую классификацию и итоговое значение метрики ROC-AUC было равно 0.75. Нормализованная матрица ошибок представлена ниже:

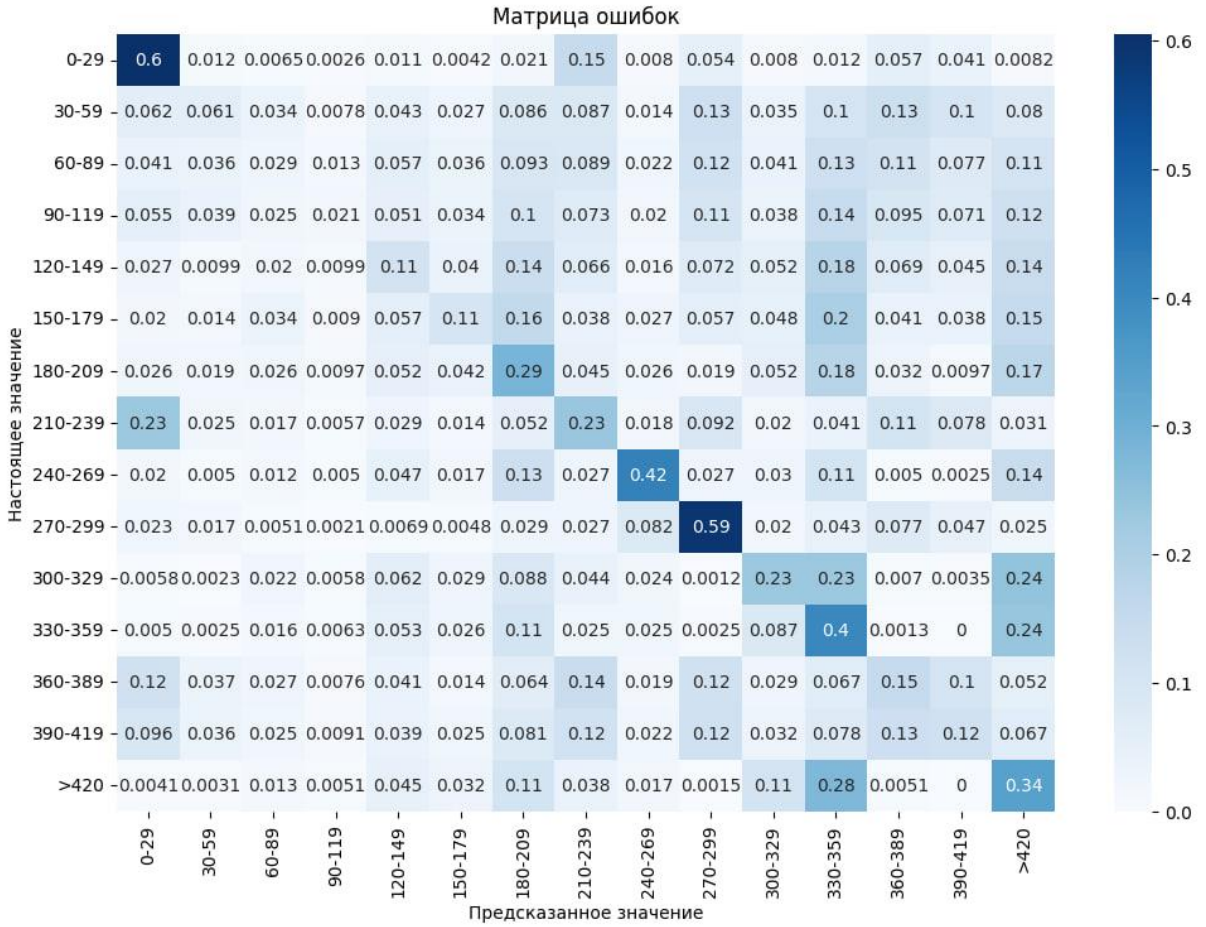


*Рисунок 8. Матрица ошибок при прогнозировании интервального значения срока экспозиции на тестовой выборке, модель 4*

В целом, результат можно считать удовлетворительным, однако, стоит отметить, что у некоторых классов (интервалов срока экспозиции) присутствует большое кол-во ошибок.

Во втором случае был получен более хороший результат по метрике MdAPE равный 0.33 на тренировочной и 0.5 на тестовой выборках.

Таким образом, приведенный анализ показывает, что прогнозирование сроков экспозиции с использованием количества просмотров в единицу времени возможно. При этом прогнозная способность моделей может быть еще несколько повышена. Следует отметить, что для обучения моделей использовалась отдельная валидационная выборка, поэтому переобучения как такового не должно было быть.

В дальнейшем результаты проведенной работы могут быть использованы для построения тепловых карт срока экспозиции объявлений по сегментам коммерческой недвижимости, в разных районах городов России, что позволит иметь возможность визуально оценить срок продажи объекта недвижимости не только в зависимости от принадлежности его к определенному сегменту и классу, но и от местоположения объекта.