4. КЛАССИФИКАЦИЯ ОБЪЕКТОВ НЕДВИЖИМОСТИ. АЛГОРИТМЫ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Исследование было проведено для следующих сегментов (подсегментов) недвижимости в разрезе потребительских классов.

4.1 КЛАССИФИКАЦИЯ ОБЪЕКТОВ НЕДВИЖИМОСТИ



1. Торговая недвижимость

— торгово-развлекательные центры и торговые центры;
— районный/микрорайонный торговый центр;
— стрит-ритейл;
— торговые площади.

Классификация, положенная в основу данного исследования, разрабатывалась в соответствие с существующими и признанными классификациями аналитических агентств, риелторов, девелоперов и консультантов, которые часто создают свои классификации для тех сегментов рынка, в которых они работают, или адаптируют западные системы стандартов к российским условиям. Отдельные варианты из них получили широкое распространение. Речь идет, например, о классификации офисной и складской недвижимости компаний Knight Frank, Swiss Realty Group, Гильдии управляющих и девелоперов, прочих компаний и отдельных авторов, таких как Стерник Г.М., а также классификация, положенную в основу справочно-методических пособий из серии «Справочник оценщика недвижимости», разрабатываемых ООО «Информ-оценка».

К объектам из рассматриваемых сегментов классификации (кроме стрит-ретейл) были отнесены как здания целиком, так и помещения, расположенные в них (используемые или планируемые к использованию по назначению в соответствие с сегментом рынка).



2. Объекты свободного назначения



3. Офисная недвижимость

— класс «А»
— класс «В»
— класс «С»



4. Производственно-складская недвижимость

— класс «А»
— класс «В»
— класс «С»

ТОРГОВАЯ НЕДВИЖИМОСТЬ

А) Торговые центры и торгово-развлекательные центры

Торговый центр — это совокупность предприятий торговли, услуг, общественного питания и развлечений, подобранных в соответствии с концепцией и осуществляющих свою деятельность в специально спланированном здании (или комплексе таковых), находящемся в профессиональном управлении и поддерживаемом в виде одной функциональной единицы.

В **торгово – развлекательных центрах** представлен комплексный спектр услуг, а также широкий выбор товаров. Особенности таких торговых центров выражены именно в широком ассортименте предлагаемых развлекательных услуг: бильярд, бани, сауны, бассейны, детские комнаты, кинотеатры.

К данному подсегменту классификации относятся окружные, региональные и суперрегиональные торговые центры.



Окружной торговый центр. Community Shopping Center.

Окружной ТЦ предлагает широкий спектр товаров и услуг, с большим выбором «мягких» товаров (мужская, женская, детская и спортивная одежда) и "жестких" товаров (металлические изделия, электроинструменты, бытовая техника). Многие центры строятся вокруг детских универмагов (junior department store), дискаунт-универмагов, крупных аптек (drug-store), универсальных магазинов, торгующих разнообразными товарами, часто по сниженным ценам (variety store), как основных арендаторов в добавлении к супермаркетам. Хотя окружной ТЦ не имеет универмага полной линии, у него могут быть сильные специализированные магазин(-ы).

► Типичный размер арендной площади – от 9500 до 47000 кв. м.

► Торговая зона в пределах 10-20 минут транспортной доступности.

► Количество посетителей – от 40 000 до 150 000 человек



Региональный торговый центр. Regional Center.

Региональные центры обеспечивают покупателей широким выбором товаров, одежды, мебели, товарами для дома (глубокий и широкий ассортимент), различными видами услуг, а также местами отдыха и развлечения. Региональные центры предлагают услуги, характерные для деловых районов, но не такие разнообразные, как в суперрегиональном торговом центре.

► Типичный размер арендной площади – от 23000 до 85000 кв. м.

► Торговая зона в пределах 30-40 минут транспортной доступности.

► Количество посетителей – от 150 000 человек



Суперрегиональный торговый центр. Super Regional Center.

Предлагает широкий выбор товаров, одежды, мебели, товаров для дома, также услуги отдыха и развлечения. Строятся вокруг 3 и более универмагов площадью не менее 7000 кв. м каждый.

► Типичный размер арендной площади – от 50000 до 150000 кв. м.

► Торговая зона может достигать 1,5 часов транспортной доступности.

► Количество посетителей – от 300 000 человек

Б) Районный/микрорайонный торговый центр

Районный ТЦ предлагает товары повседневного спроса (продукты, лекарства, хозтовары и т.д.) и услуги (прачечная, химчистка, парикмахерская, ремонт обуви) для удовлетворения ежедневных потребностей жителей района. Якорным арендатором, как правило, является супермаркет, мини-якорями могут быть аптека, хозяйственный магазин. Сопутствующими арендаторами являются магазины одежды, обуви и аксессуаров, парфюмерии, спорттоваров и пр.

► Типичный размер арендной площади – от 3000 до 10000 кв. м.

► Торговая зона в пределах 5-10 минут транспортной доступности.

► Количество посетителей – от 3 000 до 40 000 человек.

Микрорайонный ТЦ осуществляет торговлю товарами первой необходимости и предлагают услуги повседневного спроса (ремонт обуви, прачечная). Состоят как минимум из трех магазинов. Основным оператором (якорем), в большинстве случаев, является минимаркет. Примером такого рода центров могут служить бывшие советские универсамы, крупные продовольственные и промтоварные магазины, которые сократили площадь основной торговли и сдали часть площадей арендаторам (аптечный пункт, прачечная, видеопрокат, хозтовары, фотоуслуги и прочее).

► Типичный размер арендной площади – от 1500 до 3000 кв. м.

► Торговая зона в пределах 5-10 минут транспортной доступности.

► Количество посетителей – до 10 000 человек.

в) Стрит-ритейл



Street retail (стрит-ритейл) - разновидность торговой недвижимости, представляющая собой торговые помещения, расположенные на первых этажах коммерческих или жилых зданий, имеющие отдельный вход с улицы и собственные витрины. Главное требование к таким помещениям – большая проходимость.

Большая часть помещений имеет средний размер в пределах 70-150 м^2. Формат оптимально подходит для розничной торговли продовольственными и непродовольственными товарами; предприятий бьюти-сферы; финансовых и банковских организаций; бизнеса в сфере услуг; заведений общепита.

Все помещения стрит-ритейл можно разделить на две категории: элитные объекты – располагаются на оживленных улицах центральной части города, в деловом центре. Первые этажи здесь обычно заняты монобрендовыми бутиками, фешенебельными ресторанами и т.д. и спальные объекты –коммерческие площади, которые находятся на тихих улицах вблизи центра города, на оживленных магистралях периферийных районов.

Данное описание «стрит-ритейла» характерно для большинства исследуемых городов России. Однако, для таких крупных городов как Москва и Санкт-Петербург актуально понятие «торговый коридор». Под ним следует понимать улицы с наиболее активным пешеходным трафиком в связи с чем на них резко увеличивается количество объектов торговой недвижимости в формате стрит-ритейла. Среди «торговых коридоров» Москвы следует выделить следующие улицы: Кутузовский проспект, Третьяковский проезд, Старый Арбат, Новый Арбат, Столешников переулок, Петровка, Кузнецкий Мост, Никольская улица, Тверская улица на отрезке от Охотного Ряда до Пушкинской площади, Маросейка.

В Санкт-Петербурге наиболее активными улицами с большим количеством объектов торговли считаются: Средний проспект Васильевского острова (далее «ВО»), 6 и 7 линии ВО, Большой проспект Петроградской стороны, Каменноостровский проспект, Б.Конюшенная ул., основная часть Невского проспекта, Невский проспект от по. Восстания до пл. А.Невского, Владимирский проспект, ул.Рубинштейна, ул. Садовая, Московский проспект

Поэтому в данном обзоре в г. Москва и г. Санкт-Петербург проводилось независимое исследование удельных цен и ставок арендной платы класса стрит-ритейл в разрезе указанных центральных улиц, называемых «торговыми коридорами», а также прочих улиц, не относящихся к культурно- историческому центру города. Таким образом, в г. Москва и г. Санкт-Петербург, были сформированы две группы итоговых данных по объектам класса «стрит-ритейл»: «стрит-ритейл центральные улицы» и «стрит-ритейл прочие улицы».

г) Торговые площади



встроенные помещения, предназначенные для торговли, которые не располагаются в торгово-развлекательных центрах, районных/микрорайонных центрах и которые по характеристикам не могут быть отнесены к классу «стрит-ритейл». К классу «торговые площади» относятся также небольшие отдельно стоящие магазины.

ОБЪЕКТЫ СВОБОДНОГО НАЗНАЧЕНИЯ

Объекты (или помещения) свободного назначения (ПСН) — это помещения, изначально спроектированные под различные цели, без какой-либо специализации. Под ним обычно понимают отдельно стоящие небольшие здания (или же их части), которые можно использовать в иных целях. Для примера, под склад, офис, кафе или магазин. В каждом отдельном случае к помещениям свободного назначения будут предъявляться индивидуальные требования. Нередко под ПСН понимают в том числе и нежилые помещения, которые расположены на первых этажах в новых домах. Иногда ПСН встречаются и в промышленной недвижимости. Так, складское помещение может быть использовано под производство или автосервис.



ОФИСНАЯ НЕДВИЖИМОСТЬ

А) КЛАСС А



Офисные помещения класса «А» являются наиболее престижными. Офисы расположены главным образом в Центральном административном округе. Близость к метро и транспортное расположение для такого типа офисных зданий является важным критерием. Это специализированные бизнес-центры новой постройки (возраст менее 3-х лет) либо полностью реконструированные здания. Обязательно наличие полной и правильно оформленной юридической документации на право собственности и эксплуатации здания.

В офисах класса «А» высококачественное управление комплексом, отвечающее международным стандартам. Планировки открытые, позволяющие менять конфигурацию расположения офиса в соответствии с требованиями и пожеланиями арендатора, а внутренняя отделка выполняется либо по авторскому дизайну (с учетом предпочтений крупнейших арендаторов) либо по индивидуальному заказу арендатора (отделка «под арендатора») с применением высококачественных материалов. Современные качественные окна и хороший вид, обязательный атрибут офисов этого класса.

Здания оборудованы лучшими современными инженерными системами: хорошие системы вентиляции и кондиционирования, системы климат-контроля. Телекоммуникационные услуги предоставляются провайдером по оптико-волоконным сетям.

Здания обязательно имеют два независимых источника электроснабжения или наличие источника бесперебойного питания. Обязательно есть охраняемая парковка (чаще всего подобные офисные здания имеют как наземную, так и подземную стоянку) и круглосуточная охрана внутри здания: современные системы безопасности и автоматизированный контроль доступа на территорию и в помещения. Инфраструктура, как правило, обеспечивает любые нужды арендаторов. Есть конференц-зал, кафетерий/столовая для сотрудников, банкомат, фитнесс-центр, парикмахерская и другие сервисные службы.

Б) КЛАСС В

Чаще всего **офисные здания класса «В»** — это исторические здания после полного переоборудования и ремонта или бизнес-центры класса «А», после 5-7 лет эксплуатации. Иногда к этому классу относят постройки последних 10-15 лет, административные и промышленные здания советского времени после реконструкции.



На рынке офисных зданий именно этот класс бизнес-центров считается наиболее востребованным. Отличия офисов класса «В+» и «В» от класса «А» носят непринципиальный характер. Необходимо соблюдение не менее половины критериев, предусмотренных для офисных зданий класса «А». При этом обходятся офисы класса «В» арендаторам намного дешевле. Для офисов класса «В+», как правило, характерно удобное местоположение. То есть, доступность основных магистралей и близость станций метро, ухоженные территории вокруг и приятный вид из окна, внешний вид здания выглядит респектабельно, но, в отличие от класса «А», не имеет изысков в архитектуре.

Другими положительными моментами являются: грамотно оформленная юридическая документация на право собственности и эксплуатации здания, отделка офисов, фойе и лифтовых холлов на высоком уровне, но «скромнее», чем в офисных зданиях класса «А», свободная или коридорно-кабинетная планировка.

Немаловажно, что в офисах таких бизнес-центров используются функциональные системы вентиляции с предварительным охлаждением воздуха и предусмотрена возможность использования сплит-систем кондиционирования. Стоят кондиционеры относительно дешево, а качество их неоспоримо.

Владелец или управляющая компания обеспечивает оптико-волоконную связь и заключает договоры с известными телекоммуникационными провайдерами, обеспечивает круглосуточную охрану наземной парковки и здания современными системами безопасности. Как правило, в инфраструктуру бизнес-центра входит только кафетерий/столовая для сотрудников, услуги центральной рецепции в холле первого этажа и другие услуги, удовлетворяющие бытовые потребности арендаторов.

В) КЛАСС С

Офисы класса «С». К этому классу офисной недвижимости относятся, прежде всего, здания советского времени постройки (НИИ, заводы и т. д.), перепрофилированные под офисное использование. К данному классу следует так же относить качественные офисные объекты в административных и производственных зданиях, и здания, реконструированные, переоборудованные из государственных учреждений, детских садов, школ и т.п.



Офисы данного класса, как правило, обладают следующим набором характеристик:
- как правило, находятся в спальных районах города (удаленность от метро составляет чаще всего 2-3 остановки транспортом);
- фасад здания непрезентабелен и чаще всего требует косметического ремонта;
- юридическая документация на право собственности и эксплуатации здания обычно оформлена надлежащим образом;
- коридорно-кабинетная система планировки.

Для таких зданий характерно обилие несущих стен и коридоров. Внутренняя отделка выполняется преимущественно самим арендатором, а кондиционирование может вообще отсутствовать. Телекоммуникационные услуги ограничиваются телефоном и выходом в Интернет, предоставляемыми неизвестными коммерческими провайдерами. Цивилизованной парковки чаще всего нет: парковка образуется стихийно или вообще отсутствует. Управление зданием осуществляется, как правило, непосредственно силами владельца, который позаботился о круглосуточной охране объекта, но мог сэкономить на системе электроснабжения и развитой инфраструктуре.

ПРОИЗВОДСТВЕННО-СКЛАДСКАЯ НЕДВИЖИМОСТЬ

А) КЛАСС «А»

** К объектам классов А (В) относятся также и объекты промежуточных классов А+, А- (В+,В-) соответственно*

Одноэтажное складское здание из металлоконструкций и сэндвич-панелей прямоугольной формы, без колонн или с шагом не менее 12 (9) м. Расстояние между пролетами — 24 (24 м — для класса А), площадь застройки 40–45 (45-55 для класса А)%, бетонный пол с антипылевым покрытием, нагрузка не менее 5 (5) тонн на 1 кв. м, на уровне 1,20 м (1,20 м — для класса А) от земли, потолки — от 13 м (10 м — для класса А); возможность установки стеллажей, температурный режим регулируется, пожарная сигнализация, система пожаротушения, вентиляция, охрана, видеонаблюдение, электроподстанция, тепловой узел, много автоматических ворот докового типа (dock shelters) с погрузочно-разгрузочными площадками регулируемой высоты — dock levelers не менее 1 на 500 кв. м (не менее 1 на 1000 кв. м — для класса А), площадки для отстоя большегрузных автомобилей и парковки легковых машин, наличие площадок для маневрирования большегрузных автомобилей, наличие офисных помещений на территории, оптоволоконные телекоммуникации. для класса А огороженная и круглосуточно охраняемая, освещенная, благоустроенная территория, расположение вблизи центральных магистралей, профессиональная система управления, опытный девелопер, наличие системы учета и контроля доступа сотрудников, ЖД ветка

Б) КЛАСС «В»

Одноэтажное складское здание прямоугольной формы, вновь построенное или реконструированное, площадь застройки 45–55%, бетонный пол, антипылевое покрытие (асфальт или бетон — для класса В), нагрузка — 5 т/кв. м, на уровне 1,20 м от земли, потолки от 8 м (6 м для класса В), температурный режим регулируется, вентиляция, температурный режим регулируется пожарная сигнализация, система пожаротушения, вентиляция, охрана, видеонаблюдение, электроподстанция, тепловой узел, много автоматических ворот докового типа (dock shelters) с погрузочно-разгрузочными площадками регулируемой высоты — dock levelers не менее 1 на 1000 кв. м (не менее 1 на 2000 кв. м для класса В. В случае многоуровневого склада – грузовые лифты), площадки для отстоя и маневрирования большегрузных автомобилей, офисные помещения при складе, вспомогательные помещения при складе — туалеты, душевые, подсобные помещения, раздевалки для персонала, оптоволоконные телекоммуникации, огороженная и круглосуточно охраняемая, освещенная, благоустроенная территория, расположение вблизи центральных магистралей, профессиональная система управления, опытный девелопер, наличие системы учета и контроля доступа сотрудников, электроподстанция и тепловой узел, ЖД ветка. Для класса В пандус для разгрузки, охрана по периметру территории

В) КЛАСС «С»

Капитальное производственно-складское помещение или утепленный ангар, потолки — от 4 м, пол – асфальт или бетонная плитка, бетон без покрытия. В случае многоэтажного строения – наличие грузовых лифтов/подъемников, ворота на нуле, площадки для отстоя и маневрирования большегрузных автомобилей, система вентиляции, отопления, офисы на складе, пожарная сигнализация и система пожаротушения, пандус для разгрузки автотранспорта, охрана по периметру территории, телекоммуникации, вспомогательные помещения при складе, ЖД- ветка



4.2 МЕТОДИКА ОБРАБОТКИ ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ ОБЪЕКТОВ

Анализируя тексты объявлений, была сформирована база данных основных ценообразующих факторов, оказывающих непосредственное влияние на формирование стоимости и ликвидность объектов недвижимости. Подробнее с этими факторами, а также методикой работы с текстами объявлений можно ознакомится в разделе сайта «Методика и программная реализация анализа текстов объявлений».

Имея базу данных, задачу идентификации класса объекта на основе текстового описания можно свести к задаче классификации на несколько классов. Используя процедуру классификации, мы разделяем объявления по различным сегментам (подсегментам, классам, типам). Тестовой выборкой в нашем случае является база объявлений объектов недвижимости 45 городов, перечень которых представлен в основном томе Обзора, с подробным описанием.

В нашем случае для обучения модели использовалась собранная с сайтов о продаже недвижимости информация. Для текстовой классификации в данной работе были испробованы несколько методов:

- наивный байесовский классификатор,
- метод опорных векторов,
- метод k ближайших соседей,
- алгоритм случайного леса,
- градиентный бустинг,
- свёрточная нейронная сеть.

Наилучшие результаты показала классификация с использованием **градиентного бустинга**, поэтому именно этот алгоритм был применен для обучения итоговой модели и дальнейшей классификации.

При этом качество классификации можно оценивать следующими параметрами:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

где TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative. Accuracy можно интерпретировать как долю правильных ответов алгоритма, precision как долю верных срабатываний на конкретном классе, а recall как полноту найденных объектов классов. Более подробно метод и полученные результаты изложены в публикации «Анализ покупательского спроса на рынке недвижимости. индикаторы спроса и методы их определения»[1].

[1] Лейфер Л.А. Анализ покупательского спроса на рынке недвижимости. индикаторы спроса и методы их определения / Лейфер Л.А., Черная Е.В.// Вопросы оценки, №02 (92) 2018 г.

МЕТОДИКА КЛАССИФИКАЦИИ ОБЪЕКТОВ, ОСНОВАННАЯ НА ТЕХНОЛОГИЯХ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТЕКСТА ОБЪЯВЛЕНИЙ С ПРЕДЛОЖЕНИЯМИ О ПРОДАЖЕ ОБЪЕКТОВ НЕДВИЖИМОСТИ

Общее представление

Классификация объектов недвижимости исходя из анализа текста объявлений с предложениями об их продаже либо аренде состоит из следующих шагов.

1) Предобработка текста

1.1 Приведение текста к **нижнему регистру**.
1.2 Удаление из текста **стоп слов**[2], знаков препинания и чисел.
1.3 **Лемматизация**[3] (приведение слова к нормальной (словарной) форме) или **Стемминг**[4] (нахождения основы слова).
1.4 **Токенизация**[5] текста (разбиение текста на отдельные слова).

2) Векторизация текста

Существуют различные способы представления текст в векторном пространстве
Например: BagOfWords, TF-IDF, Word2Vec и т.д.

В нашей задаче применилась **TF-IDF векторизация** (подробнее см. в детальном описании, пункт.1) в сочетании с **n-грамм**.
n-грамм – это последовательность слов, которых будем рассматривать при векторизации (т.е. рассматривать будем не отдельные слова по одному, а N – заданных слов в последовательности).
n-грамм можно задать также на уровне символов.

Пример:
Есть текст - ''Продаётся коммерческое помещение в нежилом фонде''

1-грамм – *["Продаётся", "коммерческое", помещение","в","нежилом", "фонде"]*

2-Грамм – *["Продаётся коммерческое", "коммерческое помещение", "помещение в", "в нежилом", "нежилом фонде"]*

Кроме этого, необходимо задать количество признаков, так как обычно представление текста в векторном пространстве «слишком большое». Например, текст в векторном пространстве может соответствовать вектору размерности 140 тыс., то есть содержит 140 тыс. признаков.

---

[2] **Стоп-слова** – слова, которые не несут никакой смысловой нагрузки. Их еще называют шумовыми словами. Стоп-слова уже давно применяются в алгоритмах поисковых машин. Например, в английском языке — это артикли, в русском — междометия, союзы, маты и т.д.
[3] **Лемматизация** (англ. lemmatization) - это метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме (лемме)
[4] **Стемминг** (от англ. stemming) – это поиск основы слова, учитывающий морфологию исходного слова. Стемминг выполняет морфологический разбор слова, находит общую для всех его грамматических форм основу, отсекая суффиксы и окончания
[5] **Токенизация** - разбиение текста на осмысленные элементы (слова, фразы, символы), называемые токенами

3) Построение классификатора

После удачной векторизации текста можно выбрать более подходящую модель классификации и обучать. В нашей задаче использовались модель **логистической регрессии (LogisticRegression)** и модель **XGBoostClassifier.**

3.1). Модель **логистической регрессии (LogisticRegression)** сходится гораздо быстрее и интерпретируема, можно наблюдать вклад признаков при определении классов

3.2). Модель **XGBoostClassifier** по сравнению моделью логистической регрессии обучается медленнее и сложно интерпретируема, но качество классификации оказывается значительно выше.

Детальное описание

1) TF-IDF векторизация (***TF*** *— term frequency,* ***IDF*** *— inverse document frequency*)

$$\boldsymbol{tf-idf}\ (\boldsymbol{t},\boldsymbol{d},\boldsymbol{D}) = \boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) \times \boldsymbol{idf}(\boldsymbol{t},\boldsymbol{D})$$

$\boldsymbol{tf}$ – это частотность конкретного слова (n-грамма, если задано), в тексте (в одном описании)

$$\boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) = \frac{\boldsymbol{n_t}}{\sum_k \boldsymbol{n_k}}$$

где $\mathbf{n_t}$ есть число вхождений слова в тексте, а в знаменателе — общее число слов в данном тексте, **IDF** (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в текстах коллекции (во всех описании).

$$\boldsymbol{idf}\ (\boldsymbol{t},\boldsymbol{D}) = \boldsymbol{log}\frac{|\boldsymbol{D}|}{|\{\boldsymbol{d_i} \in \boldsymbol{D} | \boldsymbol{t} \in \boldsymbol{d_i}\}|}$$

где $|\boldsymbol{D}|$ — число текстов в **коллекции**[6], $|\{\boldsymbol{d_i} \in \boldsymbol{D} | \boldsymbol{t} \in \boldsymbol{d_i}\}|$ – число текстов из коллекции , в которых встречается конкретнее слово t (когда $\mathbf{n_t} \neq 0$).

Пример:

если слово **“офис”** встречается во всех описаниях, то **idf** для него получается **log(1) = 0,** а если слово “**класс а”** встречается в некоторых описаниях, то **idf** будет log(a, a>1)>1

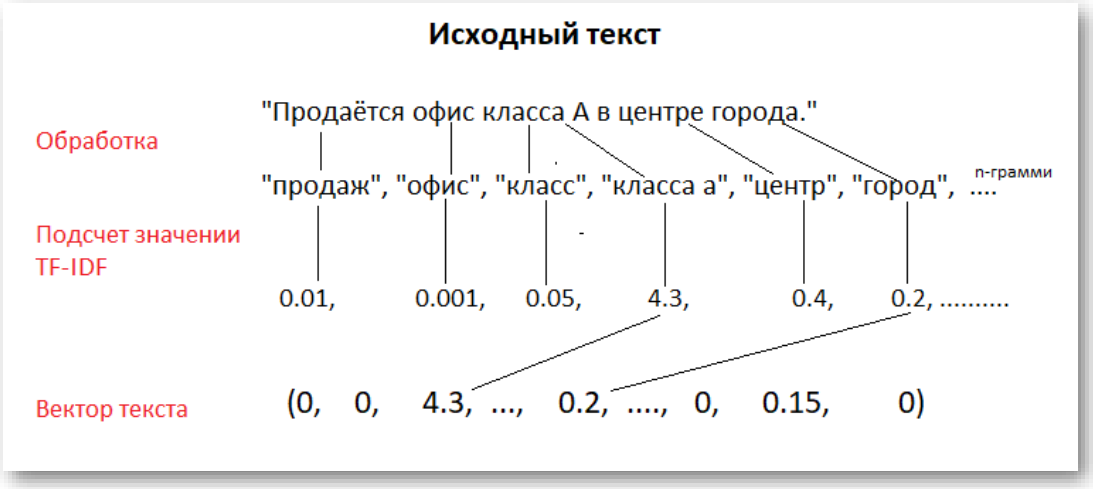
Таким образом, мера **TF-IDF** является произведением двух сомножителей:

$$\boldsymbol{tf-idf}\ (\boldsymbol{t},\boldsymbol{d},\boldsymbol{D}) = \boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) \times \boldsymbol{idf}\ (\boldsymbol{t},\boldsymbol{D})$$

Большой вес в TF-IDF получат слова с высокой частотой пределах конкретного текста и с низкой частотой употреблений в других документах.

[6] **Коллекция** – это набор всех описании

Визуализация процесса обработки и векторизации



В итоге получается вектор размерности 5000 или больше по нашему усмотрению, который можно подать на вход модели в качестве объясняющих переменных.

МЕТОДИКА КЛАССИФИКАЦИИ ОБЪЕКТОВ, ОСНОВАННАЯ НА ИСПОЛЬЗОВАНИИ БАЗЫ ЗНАНИЙ (РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ)

В данном исследовании также использовался иной способ структуризации информации, содержащейся в текстах объявлений. Основу данного метода составляет **набор регулярных выражений (база знаний)**, позволяющих путем анализа регулярных выражений осуществить разделение объектов по сегментам и представить всю информацию об объектах в виде электронных таблиц. Англоязычное название этого инструмента -*Regular Expressions* или просто *RegExp*. В качестве примера приведем набор правил, который использовался в данном исследовании для выделения объектов свободного назначения.

Фрагмент набора регулярных выражений (базы знаний) определения сегментов недвижимости

Таблица 1

Сегмент	Правила определения
Объект свободного назначения	Если в Заголовке **или** в Тексте объявления есть следующие слова: *("свободного назначения", "ПСН", "универсальн", "салон", "разного назначения", "любого вида", "любой вид", "парикмахер")* Если в Тексте есть следующие слова: *("офис")* **И** *("торгов", "магазин")* Если Сегмент **Офисный И** в Тексте есть следующие слова: ("*свободного назначения", "магазин" "ПСН", "универсальн", "любой вид", "Отдельный вход", "салон", "разного назначения", "парикмахер*")

Источник: фрагмент исследования ООО Информ-оценка

Помимо задач классификации недвижимости в разрезе сегментов и подсегментов рынка, присутствовали и не менее важные задачи классификации, например, «вычленение» площадей объектов недвижимости из текстов объявлений, определение типа сделки (Продажа/Аренда) и т.д.

МЕТОДИКА КЛАССИФИКАЦИИ ОБЪЕКТОВ, ОСНОВАННАЯ НА АНАЛИЗЕ БАЗ ДАННЫХ, ПОЛУЧЕННЫХ ОТ ЭКСПЕРТОВ

В рамках проекта работа по сбору рыночных данных проводилась в том числе с привлечением экспертов-специалистов по недвижимости. Они вручную производили сбор данных объявлений о продаже объектов недвижимости, что свидетельствует о высоком качестве данных, поскольку каждое объявление было проанализировано опытным экспертом в области недвижимости и по необходимости «прозвонено» для уточнения ряда информации.

Созданная экспертами база легла в основу классификатора, построенного методами машинного обучения и послужила «обучающей» выборкой. **В данном разделе Обзора иллюстрируется часть данного исследования (по Иркутску, Пензе, Ульяновску и Челябинску) на примере офисной недвижимости.**

Для построения классификатора использовались следующие библиотеки и настройки: pandas, numpy, seaborn, matplotlib.pyplot, scikit-learn.

Процесс классификации объектов состоит из следующих шагов (Road map):

- Первичный анализ
- Предобработка данных
- Построение классификатора (Baseline)
- Построение классификатора с разделенными моделями (Split class classifier)

ПЕРВИЧНЫЙ АНАЛИЗ

Датасет (или набор данных) был собран в автоматическом режиме, при этом были удалены признаки, связанные с ценой, поскольку датасет является общим, а ценовые признаки для аренды и продажи разные. Кроме того, ценовые признаки, скорее, являются следствием, нежели причиной маркировки класса.

В процессе обработки данных пришлось удалить 5 объектов, где не были проставлены метки классов, а также столбцы с примечаниями, которые встретились лишь в одной-двух таблицах.

В итоге была получена база данных в табличной форме следующего вида:

№	Сегмент	Класс	Тип сделки (продажа/аренда)	Заголовок объявления	Ссылка	Описание (текст)	Дата создания	Дата последнего обновления	Количество просмотров	Экспертные сроки экспозиции, в днях	...	Состояние ремонта	Офисный центр (наименование)	Год постройки	Общая площадь здания, кв.м	Площадь земельного участка, кв.м	Материал стен	Высота потолков, м	Доступ	Охрана	Парковка
0	Офисная недвижимость	класс «В»	аренда	Свободное назначение (В), 130 м²	https://volgograd.cian.ru/rent/commercial/2220...	БИЗНЕС-ХОЛЛ "LOCUS":\nСвободные пространства д...	2019-11-22 00:00:00	2019-03-25 00:00:00	33	120.0	...	ремонт класса «Премиум»	LOCUS	NaN	NaN	NaN	кирпич	NaN	свободный доступ	есть	организованная парковка
1	Офисная недвижимость	класс «В»	аренда	Офис (А), 110 м²	https://volgograd.cian.ru/rent/commercial/2266...	помещение свободного назначения общей площадью...	2020-02-25 00:00:00	2020-03-25 00:00:00	15	90.0	...	ремонт класса «Премиум»	NaN	NaN	NaN	NaN	кирпич	NaN	свободный доступ	есть	организованная парковка
2	Офисная недвижимость	класс «В»	аренда	Офис (В), 187,1 м²	https://volgograd.cian.ru/rent/commercial/2113...	офисы общей площадью 187,1 м2 в Бизнес Центре ...	2019-06-29 00:00:00	2020-03-25 00:00:00	183	180.0	...	ремонт класса «Стандарт»	Адмирал-Плаза	NaN	NaN	NaN	кирпич	NaN	свободный доступ	есть	организованная парковка
3	Офисная недвижимость	класс «С»	аренда	Офис (В-), от 29 до 170 м²	https://volgograd.cian.ru/rent/commercial/1954...	Сдаются офисные помещения в Дзержинском районе...	2019-10-21 00:00:00	2020-03-25 00:00:00	180	180.0	...	ремонт класса «Эконом»	NaN	NaN	NaN	NaN	кирпич	NaN	свободный доступ	есть	организованная парковка
4																					
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2627	Офисная недвижимость	класс «А»	продажа	Офисное помещение, 60.4 м2	https://www.avito.ru/krsk/kommercheskaya_nedv...	Продам помещение для медицинского учреждения, ...	2020-10-14 00:00:00	NaN	1529.0	90.0	...	ремонт класса «Премиум»	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2628	Офисная недвижимость	класс «В»	продажа	Офисное помещение	https://www.avito.ru/krsk/kommercheskaya_nedv...	Продам помещение 21м? по Герцена (рядом улицы:...	2020-09-28 12:37:00	NaN	537.0	90.0	...	ремонт класса «Стандарт»	NaN	NaN	NaN	NaN	кирпич	NaN	NaN	NaN	NaN
2629	Офисная недвижимость	класс «В»	продажа	Коммерческое помещение 180м2, своя парковка, 2...	https://www.avito.ru/krsk/kommercheskaya_nedv...	СРОЧНО!!!Собственник!Без комиссии!Своя парковк...	24.09.2020 20:53	NaN	5991.0	90.0	...	ремонт класса «Стандарт»	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2630 rows × 33 columns

Целевая переменная

Целевой переменной является метка класса офисной недвижимости: класс «А», класс «В», класс «С», где класс «С» имеет наименьшую стоимость.

Далее смотрелось распределение целевой переменной в данном датасете (сколько объектов принадлежат к каждому из классов).

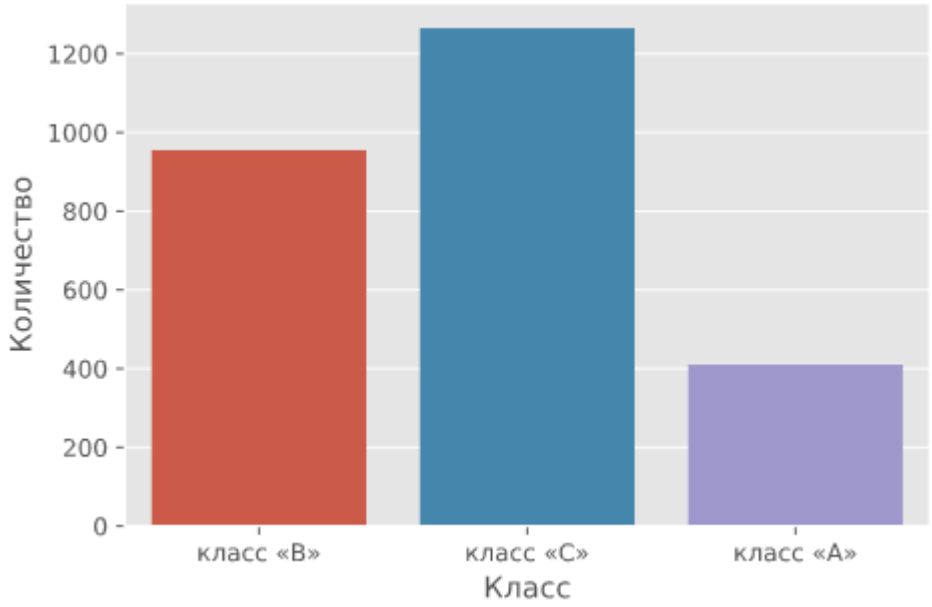


Рисунок 1. Распределение метки «класс «С» для выборки из объектов
с наименьшей стоимостью (количество меток класса «С»: 1265,
меток класса «В»: 955, меток класса «А»: 410)
Источник: Исследования ООО «Информ-оценка»

Признаки для определения целевой переменной

После для каждого из признаков датасета проверялось количество непустых значений. Оказалось, что пропусков в данных достаточно много, в некоторых признаках количество пропусков более 50% от общего количества объектов.

Часть признаков, конечно, не оказывают влияние на предсказание класса офисной недвижимости, но пропуски встречаются и в очень важных для определения класса факторах (например, в «Расстоянии до метро»).

Далее от части признаков необходимо было отказаться, а пропуски заполнить по различным стратегиям.

ПРЕДОБРАБОТКА ДАННЫХ

Удаление признаков (срез датасета)

На данном этапе формируется основной датасет, с которым продолжается дальнейшая работа. Как было указано ранее, часть признаков удаляется, таким образом основной датасет включает в себя следующие признаки: 'Класс', 'Тип здания', 'Территориальная зона', 'Планировка', 'Этаж', 'Состояние ремонта', 'Материал стен', 'Доступ', 'Охрана', 'Парковка', 'Общая площадь объекта, кв.м', 'Этажность'.

Заполнение пропусков

Пропуски в данных заполнялись постепенно, начиная с признаков с минимальным числом пустот, далее в порядке увеличения их количества. Поэтому следующим шагом выяснялось количество пропусков для каждого из признаков:

Признак	Количество пустых значений
Класс	0
Тип здания	1
Территориальная зона	1
Планировка	10
Этаж	13
Состояние ремонта	5
Материал стен	1226
Доступ	710
Охрана	929
Парковка	523
Общая площадь объекта, кв.м	0
Этажность	43

Для признаков с небольшим числом пропусков, а именно 'Тип здания', 'Территориальная зона', 'Планировка', 'Этаж', 'Состояние ремонта', пустые места заполняются как наиболее часто встречающее значение признака.

Для предсказания пустых значений признаков с большим числом пропусков строятся модели с использованием библиотек: scikit-learn (GridSearchCV, train_test_split) и catboost (CatBoostClassifier, Pool), где в качества Y принимается предсказываемый признак, а в качестве X – полностью заполненные признаки, в т.ч. те которые были заполнены на предыдущем шаге.

- Выборка делится на тренировочную (70%) и проверочную выборку (30%).
- Далее создается классификатор и задаются параметры модели.
- Обучаемся. Обучение на тренировочной выборке происходит на основе открытой программной библиотеки градиентного бустинга на деревьях решений, разработанной компанией Яндекс, CatBoost.
- Делаются предсказания и заполняются пропуски.
- Проверка качества на нескольких параметрах модели.

Таким образом, заполняются признаки: 'Парковка', 'Доступ', 'Охрана', 'Материал стен'.

Исправление значений признака 'Этажность' и заполнение пропусков 'Этажность'

Для признака 'Этажность' данные приводились к типу integer. Необходимо отметить, что предсказания категориальной модели показали плохое качество (рис. 2), поэтому в данном случае использовалась регрессионная модель с функцией потерь MAE и после обучения предсказанные значения заполнялись в пропуски.

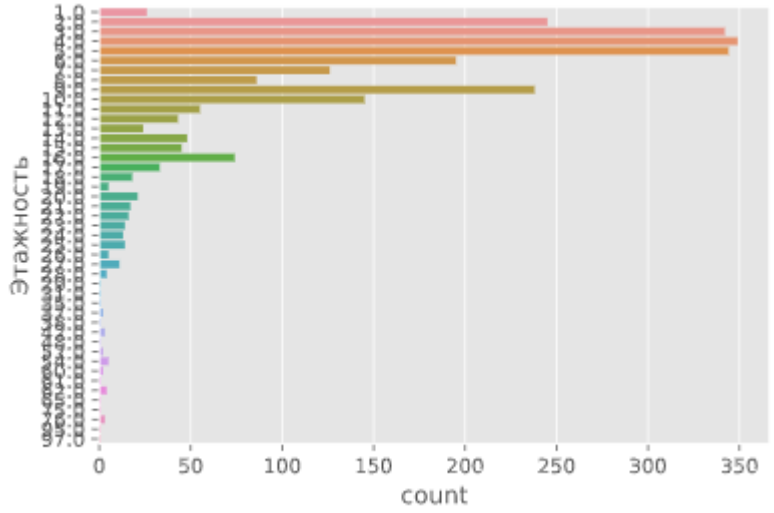


Рисунок 2. Распределение значений признака «Этажность» в выборке при использовании категориальной модели

Источник: Исследования ООО «Информ-оценка»

Исправление значений признака 'Общая площадь объекта, кв.м'

Значения признака общая площадь приводилась к единому типу данных с плавающей запятой.

Изменение признаков

Часть значений признаков необходимо привести к единой классификации, используемой в Исследовании, удалив отдельные строки из датасета, либо переименовав значения.

Парковка

Оставляем только те строки, в которых парковка принимает одно из следующих значений: 'организованная парковка', 'стихийная парковка', 'парковочных мест нет'.

До		**После**	
организованная парковка	1413	организованная парковка	1413
стихийная парковка	1147	стихийная парковка	1147
парковочных мест нет	67	парковочных мест нет	67
есть	2		
Высокий класс объясняется расположением в престижном бизнес-центре и высокой логистическй доступностью. Цена снижена за счёт уровня ремонта (стандарт)	1		

Территориальная зона

Названия территориальных зон приводится к единому виду в соответствии с правилами, приведенными в таблице ниже:

Строка содержит следующие слова/фразы	Присваиваемое наименование территориальной зоны
• Исторический центр • Центр города	Исторический центр
• Мкр • ЖК • застройк	Высотная жилая застройка
• строка содержит слово 'Центр' и при этом не содержит слова: 'города' и 'исторический'	Центр района
• строка содержит слова: 'Центр' и 'район'	Центр деловой активности
• Окраины • ромзон • краин	Окраины и промзоны
• автомагистралей	Зона автомагистралей
• ИЖС • Индивидуальные жилые дома	ИЖС
• Многофункциональные зоны	Многофункциональные зоны
• «Зеленые» зоны	Зеленая зона

До		После	
Центр города	377	Исторический центр	707
Высотная (многоквартирная) жилая застройка	371	Высотная жилая застройка	627
Центры административных районов города, зоны точечной застройки	333	Центр деловой активности	546
Исторический центр города	324	Окраины и промзоны	369
Зоны автомагистралей	252	Зона автомагистралей	257
Высотная жилая застройка	203	ИЖС	111
Центры деловой активности	162	Многофункциональные зоны	8
Окраины города, промзоны	156	Зеленая зона	2
Промзоны	115		
Индивидуальные жилые дома	100		
Окраины города	74		
Зона высотной жилой застройки	32		
Центр административного района	13		
Высотная жилая застройка 31 (Центр - Республики)	8		
Окраины	8		
Многофункциональные зоны	8		
Центр деловой активности	8		
Окраины и промзоны	7		
Окраины и промзоны 17 (Промзона ТЭЦ-1)	4		
Высотная жилая застройка 32 (Центр - КПД)	4		
Высотная жилая застройка 3 (Центр - Малыгина)	4		
Высотная жилая застройка 1 (4-й мкр., 5-й мкр., 6-й мкр.)	3		
Центр района ул. Республики 1	3		
Окраины и промзоны 16 (Промзона Электрон-Чекистов)	3		
Центр города	3		
Высотная жилая застройка 17 (Центр: 50 лет Октября)	3		

До	После
ИЖС (Коттеджная застройка Зарека) 2 Высотная жилая застройка 2 (1-й Заречный, 2-й Заречный, 3-й Заречный, Европейский) 2 ИЖС 43 (п.Дербыши, п.Московский) 2 ИЖС 4 (Червишевский тракт) 2 Окраины и промзоны 4 (Промзона Ватутина-СМП) 2 Высотная жилая застройка 10 (Дом Обороны) 2 «Зеленые» зоны 2 Центр административного района/Зона автомагистралей 2 Высотная жилая застройка 28 (1-й мкр., 2-й мкр., 3-й мкр.) 2 Высотная жилая застройка 21 (Тюменский-1, Тюменский-2, Тюменский-3, мкр. Мжк) 2 Высотная жилая застройка 31 (Центр - Республики), Зона автомагистралей 50 лет Октября 2 Центр района Воровского 2 Центр делового района 2 Высотная жилая застройка 14 (Мыс) 1 Центр района ул. Республики 2 1 Высотная жилая застройка 18 ул. Полевая 1 Окраины и промзоны 16(Промзона Электрон-Чекистов), Зона автомагистралей 50 лет Октября 1 ИЖС 9 (Озеро Цимлянское) 1 Высотная жилая застройка 27 (Воровского) 1 Центр района КПД 1 ИЖС 26 (Маяк) 1 Высотная жилая застройка 22 (Белинского) 1 Высотная жилая застройка 9 (Дом Обороны 5-ти этажные жилые дома) 1 ЖК 14 (Республика) 1 Центр района КПД, Зона автомагистралей 50 лет Октября 1 Центр района: ТРЦ Гудвин ТЦ Магеллан 1 Высотная жилая застройка 4 (Ожогина) 1 ИЖС 31 (Ватутина) 1 ИЖС 32 (Городище) 1 Центр района Исторический центр 1 Зона автомагистралей (ул. Шахтеров, ул. Енисейская, ул. Игарская) 1 ИЖС 29 (п.Метелёво) 1 Высотная жилая застройка 5 (Преображенский, Ямальский-2) 1 Высотная жилая застройка 16 (Центр) 1 Высотная жилая застройка 8 (Червишевский тракт, Южный мкр.) 1 Центр района Городище ИФНС 1 Исторический центр города, Зона автомагистралей 50 лет Октября 1 Высотная жилая застройка 13 Тарманы 1 Центр района Исторический центр 2 1	

Тип здания, планировка, территориальная зона, этаж

Удаляются строки, у которых 'Тип здания' принимает значение 'коммерческое помещение'.

До	После
офисный (бизнес) центр 1122 административное здание 657 жилой дом 598 прочие 252 коммерческое помещение 1	офисный (бизнес) центр 1122 административное здание 657 жилой дом 598 прочие 252

Удаляются строки, у которых 'Планировка' принимает значение 'встроенное помещение'.

До		После	
коридорная	1373	коридорная	1373
смешанная	650	смешанная	650
свободная (открытая)	606	свободная (открытая)	606
встроенное помещение	1		

Удаляются строки, у которых 'Этаж' принимает значения 'подвал' и 'смешанная'.

До		После	
2 и выше	1739	2 и выше	1739
1	784	1	784
цоколь	101	цоколь	101
подвал	5		
смешанная	1		

Удаляются строки, у которых 'Территориальная зона' принимает значения 'Зеленая зона' и 'Многофункциональные зоны'.

До		После	
Исторический центр	707	Исторический центр	707
Высотная жилая застройка	627	Высотная жилая застройка	627
Центр деловой активности	546	Центр деловой активности	546
Окраины и промзоны	369	Окраины и промзоны	369
Зона автомагистралей	257	Зона автомагистралей	257
ИЖС	111	ИЖС	111
Многофункциональные зоны	8		
Зеленая зона	2		

Охрана и доступ

Значения признака Охрана приводится к единому виду в соответствии с правилами, приведенными в таблице ниже:

Строка равна	Присваиваемое наименование территориальной зоны
е	есть
н	нет

До		После	
есть	1175	есть	1796
е	621	нет	834
нет	526		
н	308		

Значения признака Доступ приводится к единому виду в соответствии с правилами, приведенными в таблице ниже:

Строка равна	Присваиваемое наименование территориальной зоны
с	свободный доступ
р	расположение на закрытой территории

До		После	
свободный доступ	1562	свободный доступ	2265
с	703	расположение на закрытой территории	365
расположение на закрытой территории	358		
р	7		

Полученная база сохраняется в отдельный файл в виде:

	Класс	Тип здания	Территориальная зона	Планировка	Этаж	Состояние ремонта	Материал стен	Доступ	Охрана	Парковка	Общая площадь объекта, кв.м	Этажность
0	класс «В»	офисный (бизнес) центр	Исторический центр	свободная (открытая)	2 и выше	ремонт класса «Премиум»	кирпич	свободный доступ	есть	организованная парковка	130.0	10.0
1	класс «В»	жилой дом	Исторический центр	свободная (открытая)	1	ремонт класса «Премиум»	кирпич	свободный доступ	есть	организованная парковка	110.0	23.0
2	класс «В»	офисный (бизнес) центр	Центр деловой активности	коридорная	2 и выше	ремонт класса «Стандарт»	кирпич	свободный доступ	есть	организованная парковка	187.1	9.0
3	класс «С»	офисный (бизнес) центр	Центр деловой активности	коридорная	2 и выше	ремонт класса «Эконом»	кирпич	свободный доступ	есть	организованная парковка	41.0	4.0
4	класс «С»	административное здание	Центр деловой активности	коридорная	2 и выше	ремонт класса «Стандарт»	кирпич	свободный доступ	есть	стихийная парковка	100.0	7.0
...	...	...	...	...	...	...	...	...	...	...	...	...
2605	класс «В»	жилой дом	Высотная жилая застройка	смешанная	цоколь	ремонт класса «Стандарт»	кирпич	с	н	стихийная парковка	150.0	9.0
2606	класс «А»	жилой дом	Высотная жилая застройка	смешанная	1	ремонт класса «Премиум»	кирпич	с	е	организованная парковка	60.4	9.0
2607	класс «В»	жилой дом	Высотная жилая застройка	смешанная	1	ремонт класса «Стандарт»	кирпич	с	н	стихийная парковка	21.0	9.0
2608	класс «В»	жилой дом	Высотная жилая застройка	коридорная	цоколь	ремонт класса «Стандарт»	кирпич	с	н	стихийная парковка	180.0	9.0

2609 rows × 12 columns

ПОСТРОЕНИЕ КЛАССИФИКАТОРА BASELINE (STEP 1)

Для продолжения работы с полученной базой выборка делится на обучающую и тестовую выборки, инициализируется модель (CatBoostClassifier) с заданными параметрами, где предсказываемой величиной является Класс. Происходит обучение.

После обучения импортируются метрики оценивания моделей sklearn (precision_score, recall_score, accuracy_score, confusion_matrix). Делаются предсказания, считаются метрики и строится матрица сопряженности.

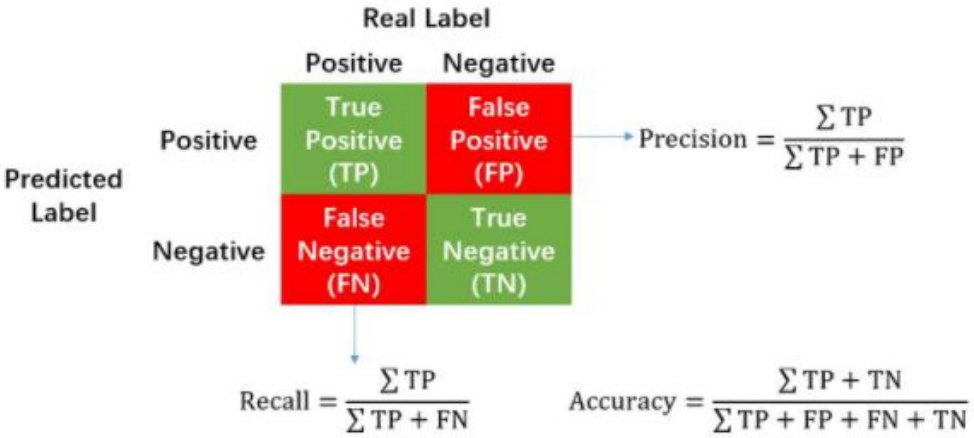


Рисунок 3. Матрица сопряженности

Источник: Исследования ООО «Информ-оценка»

Метрики Precision и Recall рассчитываются для каждого класса отдельно из-за нетривиальности расчета по всем классам вместе. Accuracy рассчитывается для всей совокупности классов.

Precision, Recall для класса "A"

```
Precision Score :  0.6730769230769231
Recall Score :  0.5555555555555556
<AxesSubplot:>
```

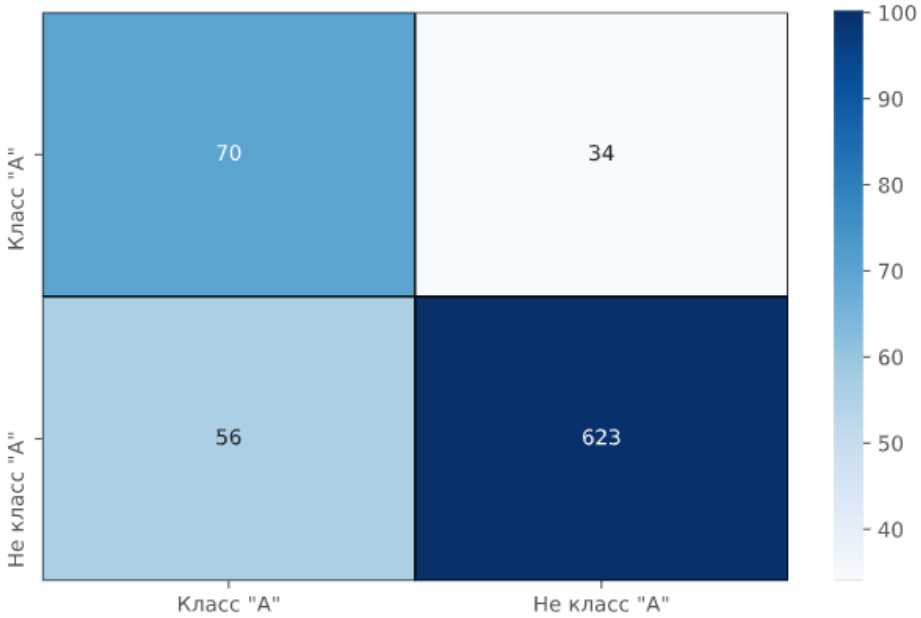


Рисунок 4. Матрица сопряженности признака «Класс «А»» в выборке

Источник: Исследования ООО «Информ-оценка»

Precision, Recall для класса "B"

```
Precision Score :  0.6377358490566037
Recall Score :  0.555921052631579
<AxesSubplot:>
```

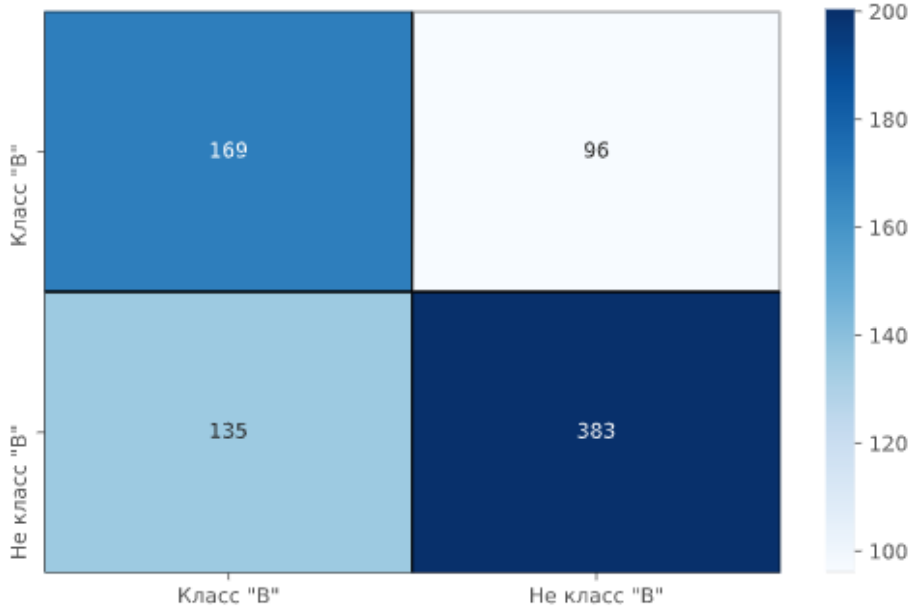


Рисунок 5. Матрица сопряженности признака «Класс «В»» в выборке

Источник: Исследования ООО «Информ-оценка»

Precision, Recall для класса "C"

```
Precision Score :  0.714975845410628
Recall Score :  0.8385269121813032
<AxesSubplot:>
```

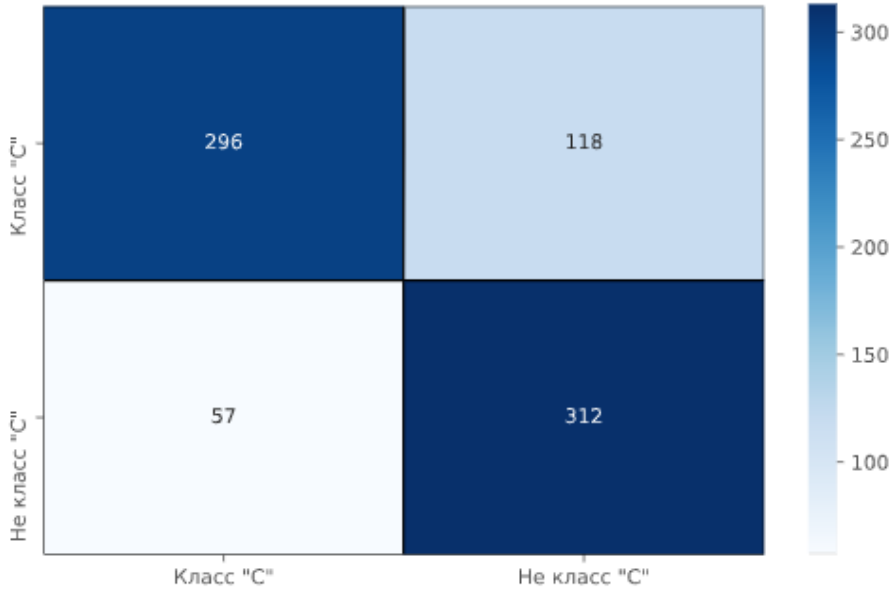


Рисунок 6. Матрица сопряженности признака «Класс «С»» в выборке

Источник: Исследования ООО «Информ-оценка»

Accuracy

```
Accuracy Score :  0.6832694763729247
<AxesSubplot:>
```

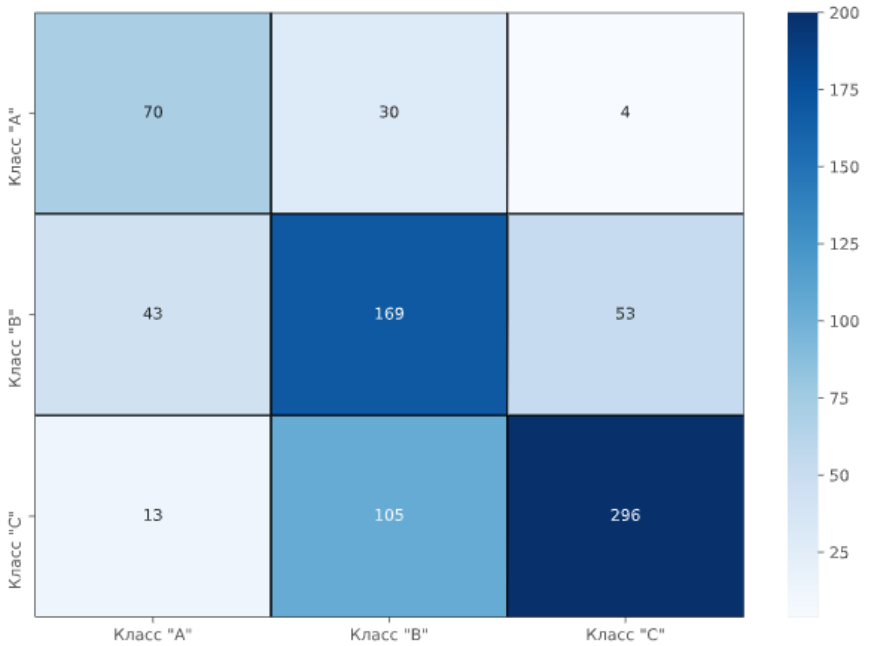


Рисунок 7. Матрица сопряженности признака «Класс» в выборке

Источник: Исследования ООО «Информ-оценка»

ПОСТРОЕНИЕ КЛАССИФИКАТОРА С РАЗДЕЛЬНЫМИ МОДЕЛЯМИ SPLIT CLASS CLASSIFIER (STEP 2)

Следующим шагом является построение классификатора с раздельными моделями. Здесь единый столбец с указанием класса в датасете разделяется на три столбца, относящихся отдельно к классу A, классу B и классу C, соответственно. Таким образом, модель для каждого из классов инициализируется своя.

	класс «A»	класс «B»	класс «C»
0	0	1	0
1	0	1	0
2	0	1	0
3	0	0	1
4	0	0	1
...	...	...	...
2604	0	0	1
2605	0	1	0
2606	1	0	0
2607	0	1	0
2608	0	1	0

2609 rows × 3 columns

Для удобства подбора параметров сперва обучаются и делаются предсказания для тех объектов, которые принадлежат к классу A, далее к классу B и в конце к классу C. Таким образом, на выходе у нас имеется три массива с предсказанными классами (y pred A, y pred B, y pred C), которые необходимо свести в единый столбец.

Обработка результатов

Создается отдельная таблица с результатами предсказаний.

	класс «A»	класс «B»	класс «C»
0	0	0	1
1	0	0	1
2	1	1	0
3	0	0	1
4	0	1	0
...	...	...	...
778	0	0	0
779	0	0	1
780	0	0	1
781	0	1	0
782	0	1	0

783 rows × 3 columns

Устанавливаются правила. Если только один классификатор присвоил объекту значение метки класса равное 1. Класс не определен, если более одного классификатора дали объекту значение метки класса 1 или если все классификаторы присвоили ему 0. Так 207 из 783 объектов попали в категорию "смешанного" класса.

В итоговую таблицу войдут только данные с «определенным» классом, данные со «смешанным» классом вырезаем. В конце объединяем обратно три столбца классов в единый столбец.

Итог

класс A

```
Precision Score :  0.84
Recall Score :  0.58
<AxesSubplot:>
```

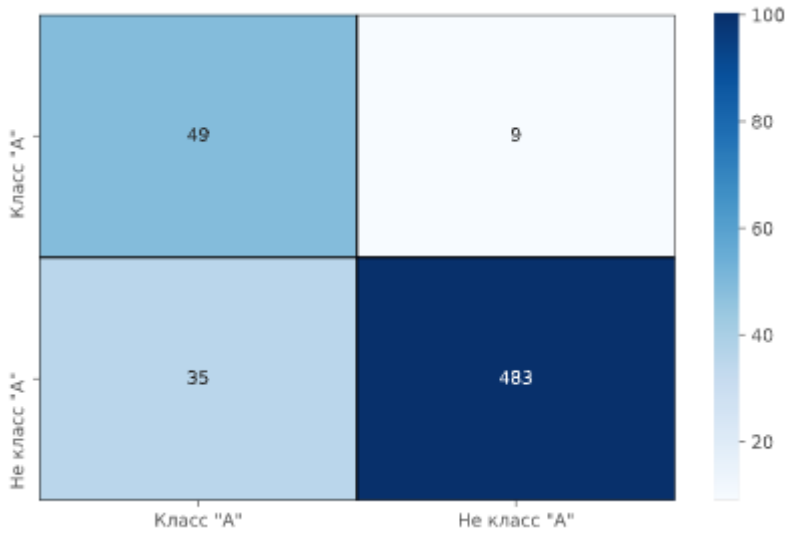


Рисунок 8. Матрица сопряженности признака «Класс «A»» в выборке

Источник: Исследования ООО «Информ-оценка»

класс B

```
Precision Score :  0.7
Recall Score :  0.59
<AxesSubplot:>
```

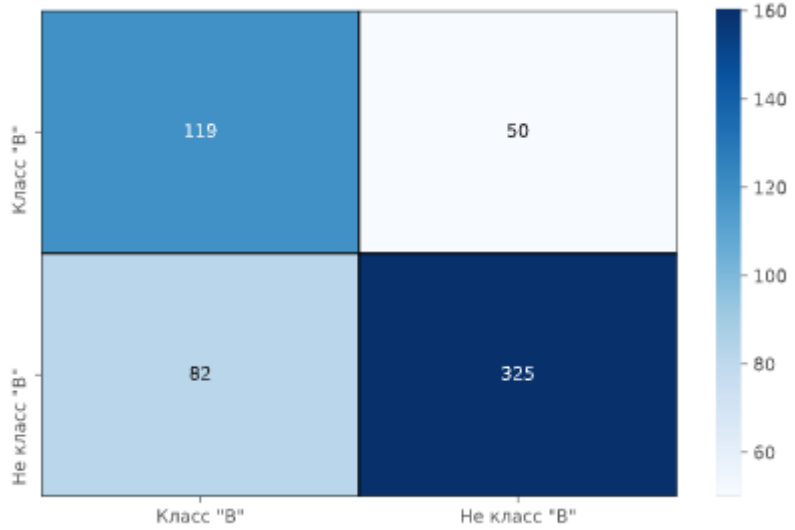


Рисунок 9. Матрица сопряженности признака «Класс «B»» в выборке

Источник: Исследования ООО «Информ-оценка»

класс С

```
Precision Score :  0.77
Recall Score :  0.92
<AxesSubplot:>
```

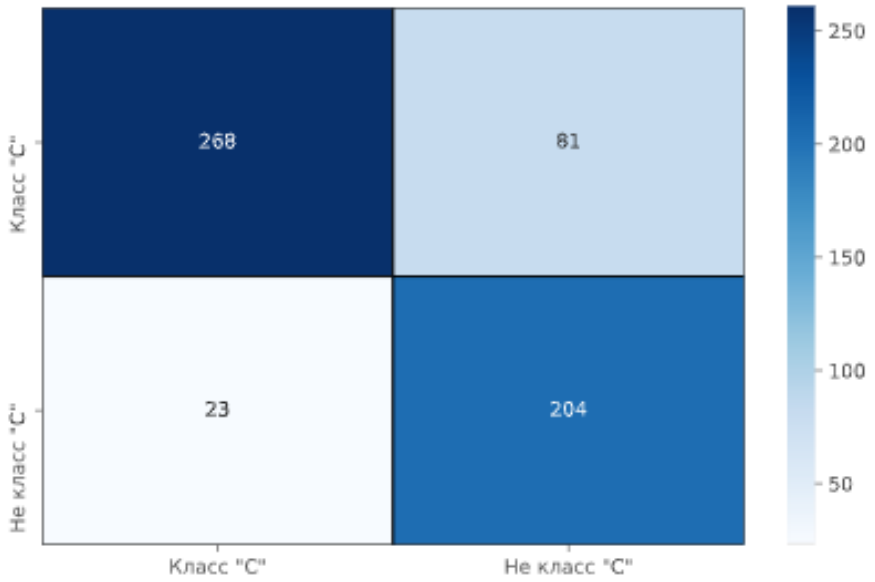


Рисунок 10. Матрица сопряженности признака «Класс «С»» в выборке

Источник: Исследования ООО «Информ-оценка»

Accuracy

```
Accuracy Score :  0.76
<AxesSubplot:>
```

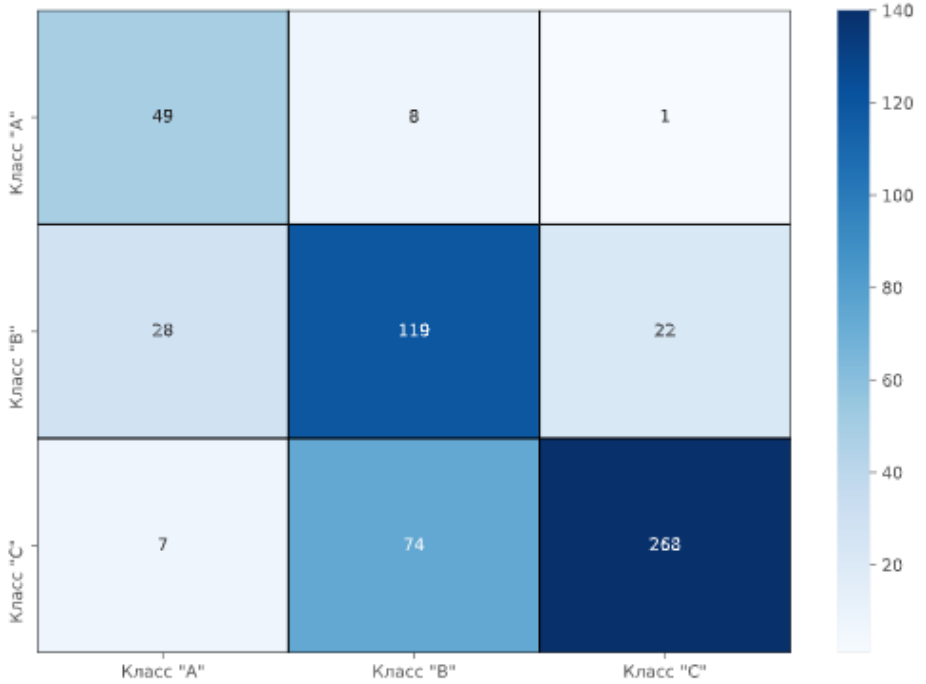


Рисунок 11. Матрица сопряженности признака «Класс объекта» в выборке

Источник: Исследования ООО «Информ-оценка»