**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет «Высшая школа бизнеса»

Департамент Бизнес-информатики

**Проектное задание по научно-исследовательскому семинару**

**“Информационная бизнес-аналитика”**

**Работу выполнили:**

Качало Павел Станиславович, 211

Подкопаев Никита Кириллович, 211

Емелин Владимир Павлович, 211

Сорокин Михаил Евгеньевич, 215

Бегеба Денис Владимирович, 212

**Москва 2024 г.**

# **Оглавление:**

[**Оглавление: 2**](#_gt13fq8rp04v)

[**1. Введение 3**](#_66v5zt95z1nd)

[1.1 Цель и задачи проекта 3](#_hu9r8e3ymmyh)

[1.2 Рамки проекта и ожидаемые результаты 4](#_qajrrawr54r6)

[**2. Основная часть 5**](#_9ykgqjqh2bmf)

[2.1 Анализ требований к решению 5](#_r6y4zk6zt5t5)

[2.2 Обзор рынка и выбор решения 7](#_jtvsp6625u85)

[2.3 Предлагаемое решение 11](#_yn468uxgltj0)

[Данные 11](#_uz1w2uf3o17e)

[Модель 13](#_3eg6zx8g4hoc)

[ТГ-бот 14](#_be8ange77766)

[2.4 Контрольный пример(демонстрация) 15](#_andb9gwxucy2)

[2.5 План реализации по вехам 15](#_obyhz0b5w0ci)

[**3. Заключение 16**](#_cz4phojbqugc)

[3.1 Заключение проекта 16](#_mrglcoqewxhb)

[3.2 Возможные риски и ожидаемые эффекты 16](#_ouqp46z3l4vx)

[**4. Библиографические источники 17**](#_ux6w80yjkdvi)

# **1. Введение**

Эффективная оценка стоимости объектов недвижимости определяет успех компаний в данной отрасли. Также наблюдается рост спроса на цифровые решения, которые позволяют оперативно и достаточно точно давать такую оценку.

Наиболее часто применяются дорогостоящие и времязатратные традиционные подходы к оценке недвижимости, основанные на экспертных оценках и стандартных методах анализа. В редких случаях человеческий фактор может снижать их точность.

Использование машинного обучения в оценке стоимости объектов недвижимости позволяет анализировать огромные массивы данных, выявлять закономерности и предоставлять результаты, которые с достаточной степенью точности отражают рыночную стоимость объектов. Это делает применение таких технологий особенно актуальным для компаний, стремящихся повысить эффективность работы.

В рамках данной работы рассматривается задача автоматизации оценки стоимости недвижимости с использованием технологий машинного обучения. Основное внимание уделяется разработке системы, которая позволяет агентству недвижимости быстро и точно определять стоимость квартир для принятия решение о последующем выкупе.

Объектом исследования в рамках данной работы выступает процесс оценки стоимости объектов недвижимости.

Предметом исследования — методы машинного обучения, применяемые для автоматизации такого процесса.

Потенциальным заказчиком выступает агентство недвижимости, которое занимается исключительно экспертной оценкой, и заинтересовано в повышении оперативности оценки стоимости недвижимости, снижении затрат, связанных с проведением экспертной оценки, и привлечении новых клиентов.

## **1.1 Цель и задачи проекта**

**Целью данной работы** является оптимизация процесса оценки стоимости объектов недвижимости, основанная на применении технологий машинного обучения.

**Задачи проекта:**

1. Изучение актуальности проблемы.
2. Изучение процесса оценки стоимости недвижимости.
3. Сбор и предварительная обработка данных для обучения модели.
4. Формирование требований для работы системы.
5. Создание моделей оценки стоимости квартиры.
6. Выбор модели с учетом специфики деятельности.
7. Разработка ТГ-бота, который будет работать с помощью нашей модели.
8. Оценка точности модели и её оптимизация.
9. Тестирование ТГ-бота и системы с реальными пользователями.

**Состав команды:**

* Качало Павел - разработка тг-бота.
* Подкопаев Никита - разработка ML-модели.
* Емелин Владимир - диаграммы, общая информация.
* Сорокин Михаил - введение, цель и задачи, BPMN-диаграммы, возможные риски и ожидаемые эффекты.
* Бегеба Денис - введение, обозначение рамок проекта, обзор рынка и выбор решения, план реализации по вехам.

## **1.2 Рамки проекта и ожидаемые результаты**

**Ограничения проекта:**

* Качество исходных данных – для обучения модели использован набор данных с платформы Kaggle, содержащий информацию о недвижимости на ноябрь 2023 года. Данные были очищены.
* Функциональность ТГ-бота – для телеграмм-ботов установлены ограничения на количество символов в сообщении, описании самого бота и его команд, также ограничена частота отправки сообщений пользователям и запросов к API. Эти ограничения необходимо учитывать при разработке.

**Допущения:**

* Статичность данных – в рамках пилота для обучения модели будет использован ранее упомянутый статичный набор данных с информацией на ноябрь 2023 года, динамическое обновление данных на этом этапе не предусмотрено.
* Расчёт стоимости – при расчёте стоимости недвижимости будут учитываться исключительно характеристики недвижимости. Не учитывается влияние внешних факторов: инфляции, ключевой ставки, спроса и т.д.
* Территориальные ограничения – на этапе пилота наш бот будет работать только в рамках Москвы и ближайшего Подмосковья.

**Ожидаемые результаты:**

* Разработан ТГ-бот, которая автоматически оценивает стоимость недвижимости на основе указанных пользователем характеристик.
* Подготовлен отчёт и презентация результатов.

# **2. Основная часть**

## **2.1 Анализ требований к решению**

**Функциональные требования**

1. ***Получение данных от пользователя.*** Система должна запрашивать информацию о квартире, такую как:
   1. Находится ли квартира в новостройке или нет
   2. В скольких минутах от метро находится квартира
   3. Находится ли квартира в Москве или в Подмосковье
   4. Сколько комнат в квартире
   5. Какова общая площадь квартиры
   6. Какова площадь кухни
   7. На каком этаже находится квартира
   8. Сколько этажей всего в доме
   9. Какой ремонт у квартиры
2. ***Обработка входных данных.*** Система должна валидировать введенные данные (например, проверка, что площадь — это число).
3. ***Оценка стоимости квартиры.*** Система должна интегрировать и использовать обученную ML модель для оценки стоимости квартиры на основании введенных данных.
4. ***Предоставление результата.*** После оценки стоимость квартиры должна быть представлена пользователю с пояснением.

**Нефункциональные требования**

1. ***Производительность*.** Система должна обрабатывать запросы и предоставлять результат не более чем за 3 секунды.
2. ***Надежность*.** Система должна быть доступна 24/7 с минимальным временем простоя.
3. ***Удобство использования***. Интерфейс взаимодействия должен быть интуитивно понятным для пользователей.
4. ***Безопасность*.** Система должна защищать личные данные пользователей и предотвращать несанкционированный доступ к информации.
5. ***Масштабируемость*.** Система должна быть способен обрабатывать увеличивающееся количество пользователей и запросов без потери производительности.
6. ***Логирование и мониторинг*.** Необходимо вести журнал запросов и ошибок для дальнейшего анализа и улучшения работы системы.

**Диаграмма Состояний**Диаграмма состояний Системы представляет собой линейный процесс, в котором пользователь проходит через последовательность шагов. Каждый шаг строго связан с предыдущим, что исключает возможность пропуска этапов или ошибок при вводе данных. Такой подход обеспечивает структурированное и удобное взаимодействие с ботом, позволяя пользователю вводить информацию об объекте недвижимости последовательно и логично.

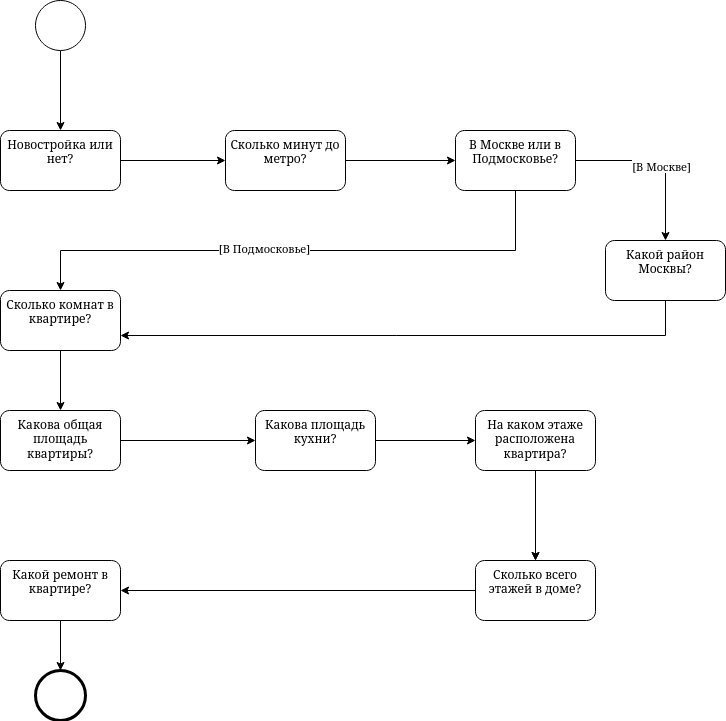


Рис №1 (Диаграмма состояний)

## **2.2 Обзор рынка и выбор решения**

**Экспертная оценка:**

Сейчас для наиболее точной оценки стоимости недвижимости используется экспертная оценка, которая проводится сертифицированным оценщиком. Для её проведения клиенту необходимо обратиться в оценочную организацию, оплатить услугу оценки, подготовить необходимые документы, договориться о дате и времени осмотра (при необходимости) и дождаться подготовки отчёта об оценке объекта.

Этот метод является самым точным, так как учитывает специфику объекта, и применяет наиболее актуальный в каждом конкретном случаем подход к расчёту рыночной стоимости: сравнительный, затратный или доходный.

К недостаткам можно отнести то, что услуга не бесплатная и требует временных затрат.

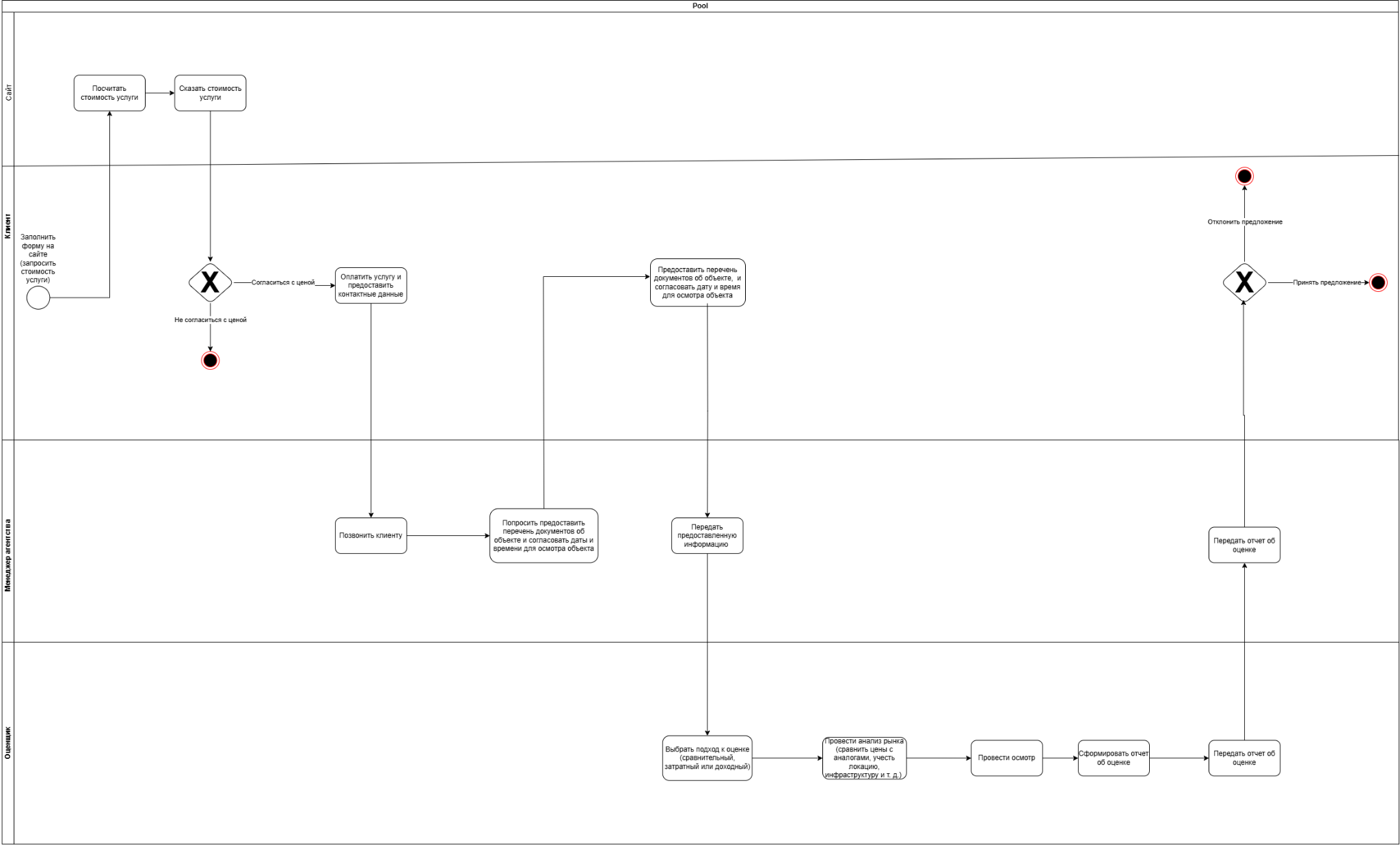


Рис №2(BPMN-диаграмма AS IS)

**Онлайн-калькуляторы и автоматизированные системы:**

Бесплатной и более быстрой альтернативой привычной экспертной оценке являются онлайн-калькуляторы и автоматизирование системы оценки стоимости недвижимости. Такие решения прогнозируют стоимость на основе информации об объекте, предоставленной пользователем, и собственных данных о схожих объектах, инфраструктуре рядом с объектом и тренда цен на рынке.

Эти решения являются более доступными и быстрыми, однако результатом их работы является лишь примерная стоимость.

**Выбор решения:**

Исходя из бизнес-потребностей потенциального заказчика и определённых ранее требований, было принято решение разработать телеграм-бот. Выбор обусловлен относительной простотой разработки и удобством для конечного пользователя.

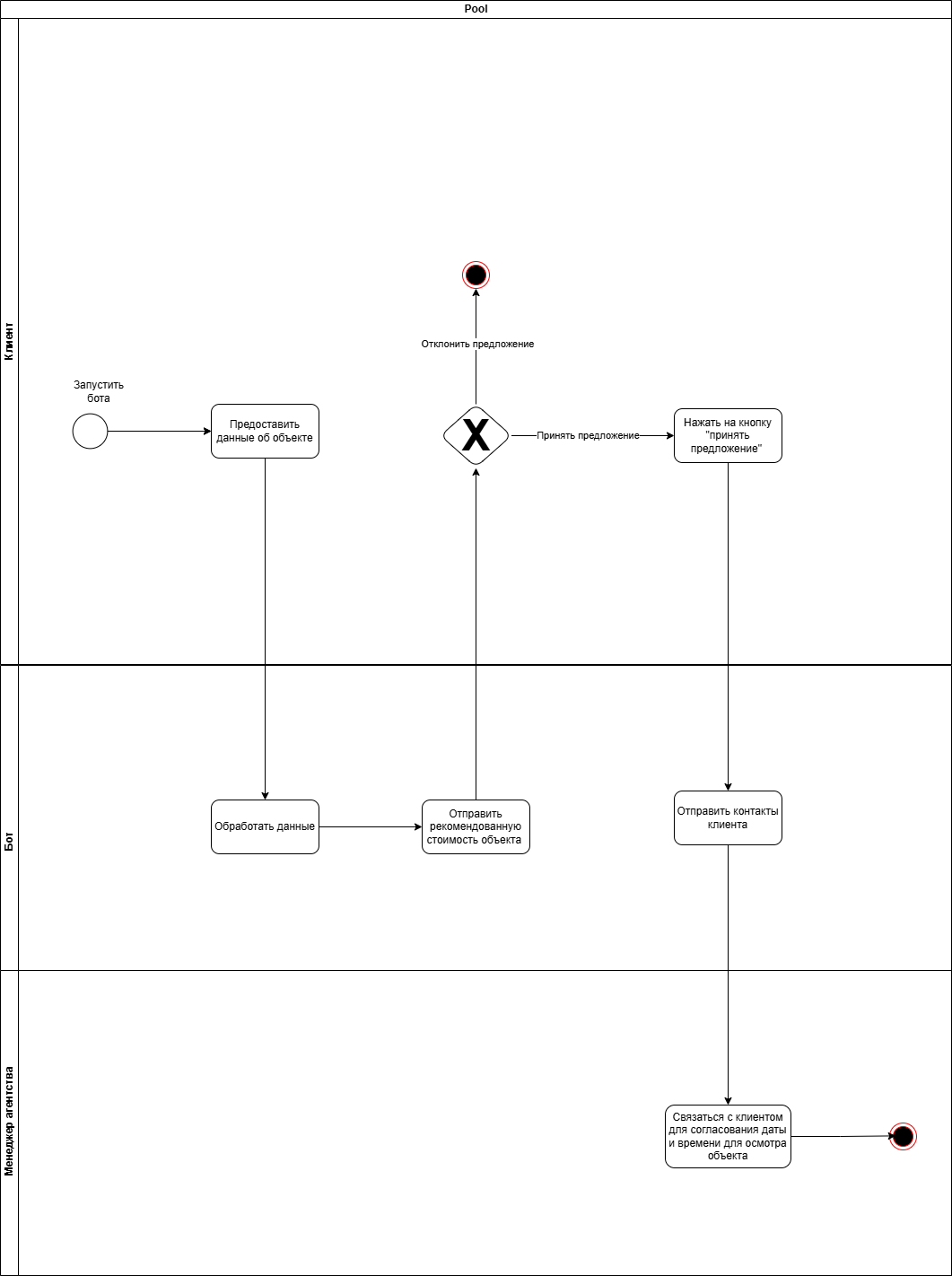


Рис №3 (BPMN-диаграмма TO BE)

## **2.3 Предлагаемое решение**

В качестве решения мы предлагаем создать ТГ-бота, который будет по входящим параметрам оценивать стоимость квартиры и выдавать ее пользователю. Данное решение глобально состоит из двух частей: модель, которая предсказывает цену и ТГ-бот, который будет работать с помощью данной модели.

### **Данные**

Однако прежде чем создавать собственно модель необходимо поработать с данными. У нас есть датасет с параметрами квартир москвы и ближнего подмосковья. Помимо стандартного EDA для улучшения качества нашего бота(избавиться от пропусков и выбросов, убрать ненужные признаки, избавиться от мультиколлинеарности) также необходимо поделить наш датасет на несколько частей с разными районами, поскольку стоимость квартиры очень сильно зависит от ее расположения(разница в стоимости между 2 квартирами одной площади, но в разных районах может достигать десятков раз). Это оказалось нетривиальной задачей, т.к. единственные показатели, связанные с расположением квартир в наших данных были метро и индекс москва или подмосковье. Для начала мы поделили наши данные на москву и подмосковье. Далее нам необходимо было поделить квартиры в москве на район, для этого мы решили создать новый признак “district” или же в нашем случае округ, который показывает в каком административном округе(далее АО) москвы находится квартира. С помощью списка станций метро в каждом АО мы смогли определить округ для каждой квартиры. Дальше мы сгруппировали все квартиры по округам и составили таблицу где находится разная информация о стоимости квартиры и в итоге получили следующую таблицу:

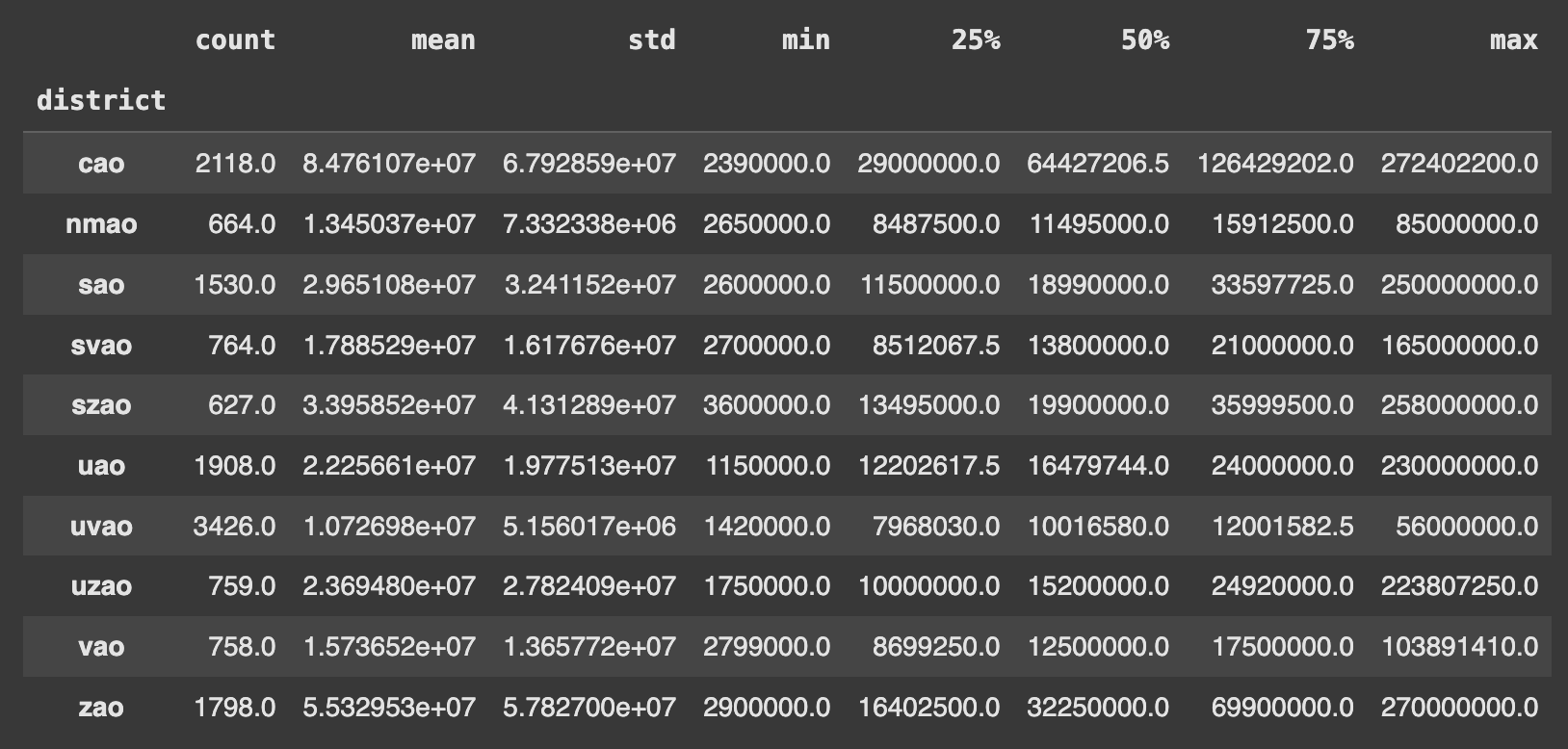


Рис №4(сравнение показателей разных АО)

Внимательно изучив данную таблицу мы решили поделить москву на 3 района:

* moscow\_high: cao, zao
* moscow\_medium: sao, szao, uao, uzao
* moscow\_low: nmao, uvao, vao, svao

В итоге мы получили такую ситуацию:

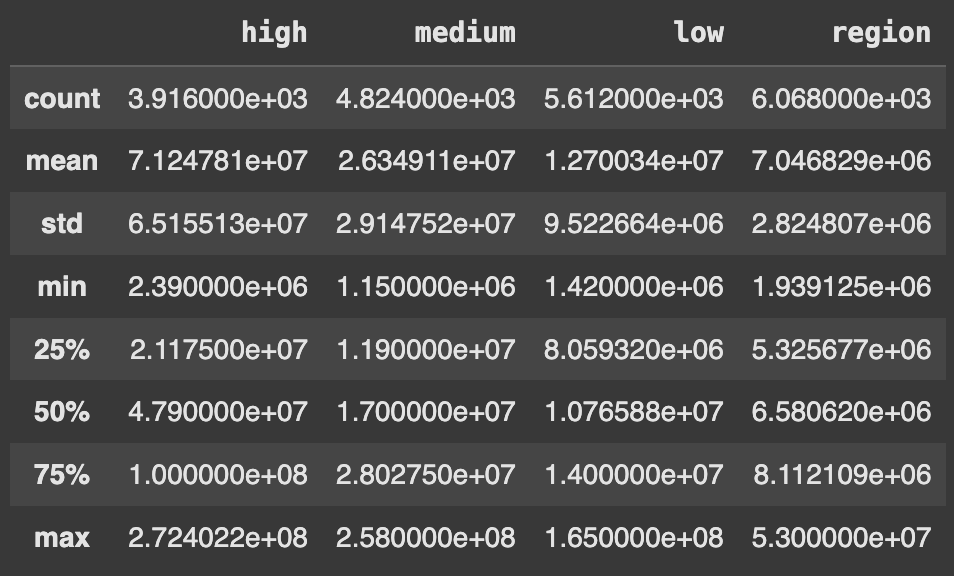


Рис №5 (показатели цены разных район)

### **Модель**

Чтобы достичь лучшего качества мы решили попробовать воспользоваться различными алгоритмами и в конце для каждого района выбрать подходящую. В качестве моделей мы решили выбрать различные имплементации градиентного бустинга(XGBoost, Lightgbm, Catboost), т.к. в настоящий момент градиентный бустинг является самым сильным алгоритмом классического ML. Также для более точного подбора гиперпараметров мы воспользовались специальной библиотекой Optuna, которая работает на основе алгоритмов TPE и CMA-ES.

В качестве функции потерь мы выбрали квантильную функцию потерь. Такой выбор мы сделали из-за того, что она может по-разному штрафовать прогноз выше и ниже целевого значения, а в нашем случае очень важно, чтобы предсказания были либо выше, либо ниже ответа, т.к. агентству по недвижимости потом еще надо будет вычитать какой-то процент из получившейся суммы. Если же модель будет и пере- и недопрогнозировать, то в таком случае неясно какой именно процент надо вычитать, чтобы агентству было выгодно.

В качестве метрик мы выбрали RMSE и . RMSE у нас выступает в качестве основной метрики. Однако еще мы решили взять , поскольку показатели RMSE не интерпретируемые, т.е. по ним можно сравнить 2 модели и понять какая из них лучше, но узнать насколько высока предсказательная способность модели в принципе, нельзя, в то время как может дать ответ на этот вопрос, поскольку она лежит на интервале от 0 до 1 и фактически показывает процент дисперсии, который наша модель смогла объяснить.

В итоге мы для каждого района составили табличку, где указано значение ошибки RMSE выраженная в миллионах, а также среднее время предсказания модели, выраженное в миллисекундах

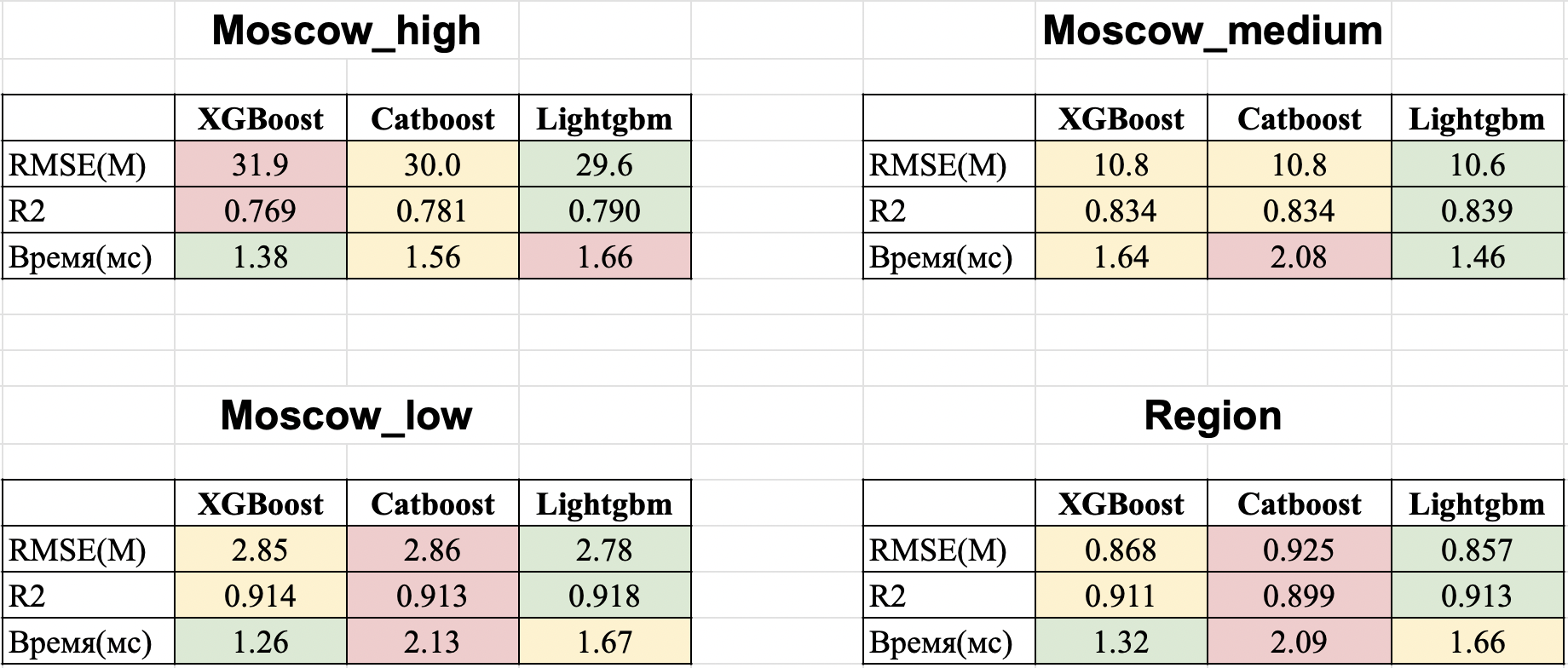


Рис №6 (сравнение моделей)

Изучив результаты мы решили выбрать модели на основе Lightgbm для всех районов, поскольку она показала лучшее качество. По части времени модель тоже показала себя неплохо, не во всех случаях она была самой быстрой, но учитывая, что время работы маленькое, то это не критично.

Более подробно ознакомиться с обучением моделей можно по ссылке: <https://colab.research.google.com/drive/1v4uYgsZ9XQ0eXO4uWo-GnvclbXNoFSRX?usp=sharing>

### 

Рис №7 (Архитектура приложения)

### **ТГ-бот**

В качестве инструмента разработки телеграм-бота, был использован язык программирования Python и официальная библиотека python-telegram-bot, которая является асинхронным интерфейсом для Telegram Bot API, разработанным командой Telegram.

Такую систему можно отнести к монолитной клиент-серверной архитектуре, где backend приложение исполняется и обрабатывает все запросы на сервере, а в качестве клиента выступает диалоговое окно с ботом в Telegram.

Взаимодействие с ботом происходит в виде последовательных вопросов про квартиру со стороны бота и ответов со стороны пользователя. Такие вопросы, как «Новая ли квартира или нет?», «Квартира находится в Подмосковье или в Москве?», «В каком районе Москвы находится квартира?», «Какой ремонт у квартиры?» предполагают выбор ответа из заранее подготовленных и представленных пользователю. Ответами для остальных вопросов являются написанные пользователем текстовые сообщения.

После получения ответов на все вопросы из файла импортируется соответствующая модель машинного обучения, после чего, используя данные о квартире, та представляет свою оценку стоимости квартиры и отправляет ее пользователю.

Ознакомиться с кодом можно в гитхаб-репозитории: <https://github.com/opavelkachalo/mos_real_estate_bot>

## **2.4 Контрольный пример(демонстрация)**

<https://disk.yandex.ru/d/SPqLUpXzg_kgjQ>

## **2.5 План реализации по вехам**

Рис №8 (План реализации)

# **3. Заключение**

## **3.1 Заключение проекта**

В рамках проекта мы провели анализ области оценки недвижимости, разобрали и сформировали функциональные и нефункциональные требования к решению, проанализировали бизнес процессы, смоделировав их с использованием нотации BPMN, составили диаграмму состояний для телеграм-бота, нашли датасет с данными о характеристиках и стоимостями квартир в Москве и Подмосковье, провели EDA с данными из этого датасета, реализовали и обучили разные модели ML в зависимости от района и уровня стоимости квартир, разработали телеграм-бота для взаимодействия с пользователями, интегрировали полученные ранее модели с телеграм-ботом для расчета стоимости квартир пользователей, сформировали план реализации проекта по вехам. В результате получили рабочий MVP, который в дальнейшем можно будет развивать.

## **3.2 Возможные риски и ожидаемые эффекты**

**Возможные риски**

1. Риски, связанные с данными: потеря актуальности данных в условиях быстро меняющегося рынка недвижимости.
2. Риски, связанные с хостинг-провайдером: возможны сетевые атаки на подсети хостинг-провайдера, что может привести к временной потере доступа к функциям ТГ-бота.
3. Зависимость от платформы Telegram: политика Telegram в отношении ботов может измениться, что повлияет на возможности использования платформы для бизнес-целей.

**Ожидаемые эффекты**

1. Оптимизация процесса оценки недвижимости.
2. Сокращение затрат.
3. Привлечение большего числа клиентов.
4. Создание конкурентного преимущества за счет внедрения технологий машинного обучения.
5. Возможность адаптации модели и бота для оценки других типов недвижимости.

# **4. Библиографические источники**

1. Инком недвижимость(2024) Срочный выкуп квартир // Сайт <https://www.incom.ru> (<https://www.incom.ru/srochny-vikup/> )
2. MWMoskva(2024) Станции метро по округам // Сайт <https://mwmoskva.ru> (<https://metro.mwmoskva.ru/stancii-spisok/po-okrugam/> )
3. Открытая Москва(2024) Административные округа Москвы // Сайт <http://mosopen.ru> (<http://mosopen.ru/regions> )
4. Хабр(2024) Это мы юзаем: библиотека Optuna в Python для оптимизации гиперпараметров // Сайт <https://habr.com/ru/feed/> (<https://habr.com/ru/companies/otus/articles/801463/> )
5. Апхилл(2024) Оценка жилой недвижимости // Сайт <https://www.uphill.ru> (<https://www.uphill.ru/uslugi/oczenka-stoimosti/oczenka-zhiloj-nedvizhimosti/> )
6. Апхилл(2016) Об оценке рыночной и ликвидационной стоимостей недвижимого имущества // Сайт <https://www.uphill.ru> ([https://www.uphill.ru/assets/examples%20of%20reports/Ocenka-zhiloi-nedvizhimosti/Отчет%20об%20оценке%20квартиры%20в%20г.%20Москве.pdf](https://www.uphill.ru/assets/examples%20of%20reports/Ocenka-zhiloi-nedvizhimosti/%D0%9E%D1%82%D1%87%D0%B5%D1%82%20%D0%BE%D0%B1%20%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B5%20%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D1%80%D1%8B%20%D0%B2%20%D0%B3.%20%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B5.pdf) )
7. Циан Методология оценки // Сайт <https://support.cian.ru/ru> (<https://support.cian.ru/ru/knowledge_base/art/300/cat/89/> )
8. python-telegram-bot(2024) Python Telegram Bot Documentation // Сайт <https://docs.python-telegram-bot.org/en/stable/index.html#>
9. Kaggle(2023) Moscow housing price dataset // Сайт <https://www.kaggle.com> (<https://www.kaggle.com/datasets/egorkainov/moscow-housing-price-dataset> )

# 