Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет "Высшая школа экономики"
Факультет экономических наук
Образовательная программа "Экономика и экономическая политика"

Домашняя работа по дисциплине:

"Машинное обучение в экономических исследованиях"

"Прогнозирование цены 1 м^2 на жилую недвижимость в Москве и МО"

Работу выполнили: студенты гр. МЭКЭП231
Журба Анна,
Козловская Полина

Преподаватель: Потанин Богдан Станиславович

Москва, 2024

1 Обоснование темы

Задание 1.1. Придумайте непрерывную зависимую (целевую) переменную (например, заработная плата или прибыль) и бинарную переменную воздействия (например, образование или факт занятий спортом).

Ответ: В качестве непрерывной зависимой переменной мы выбрали цену 1 $м^2$ жилой недвижимости в Москве и Московской области.
Бинарная переменная воздействия - наличие станции метро в шаговой доступности, то есть в 10 минутах ходьбы (1 - станция метро находится в 10 минутах ходьбы, 0 - иначе).

Задание 1.2. Опишите, для чего может быть полезно изучение влияния переменной воздействия на зависимую переменную. В частности, укажите, как эта информация может быть использована бизнесом или государственными органами.

Ответ: Изучение влияния наличия станции метро на цену квадратного метра жилой недвижимости является важным инструментом для оценки рыночной стоимости, привлечения инвестиций, улучшения качества жизни и понимания динамики рынка. Эта информация полезна как для бизнеса, так и для государственных органов, позволяя им принимать более обоснованные решения в своих сферах деятельности.
Во-первых, знание о том, как наличие метро влияет на цены на неджвижимость, помогает риелторам и застройщикам устанавливать более точные цены на жилье. Это позволяет им лучше оценивать стоимость объектов в зависимости от их расположения, что важно при продаже и аренде недвижимости.
Для государственных органов анализ цен на жилье в зависимости от транспортной доступности может помочь в планировании городской инфраструктуры и транспортной сети. Это может привести к более обоснованным инвестициям в развитие метро и других видов транспорта.
Понимание влияния транспортной инфраструктуры на стоимость недвижимости может помочь в разработке стратегий по привлечению инвестиций в новые районы, особенно те, которые находятся в процессе развития.
Кроме этого, застройщики могут использовать информацию о транспортной доступности для создания более привлекательных жилых комплексов, что может повысить качество жизни жителей и увеличить спрос на жилье.

Задание 1.3. Обоснуйте наличие причинно-следственной связи между зависимой переменной и переменной воздействия. Приведите не менее 2-х источников из научной литературы, подкрепляющих ваши предположения.

Ответ: Причины наличия причинно-следственной связи между зависимой переменной и переменной воздействия:
1. Наличие станции метро значительно увеличивает доступность района, что делает его более привлекательным для покупателей и арендаторов. Это приводит к росту спроса на жилье в таких районах, что, в свою очередь, повышает цены.

2. Близость к метро улучшает качество жизни жителей, так как позволяет быстрее и удобнее добираться до работы и других важных мест. Это также может увеличить стоимость недвижимости, поскольку покупатели готовы платить больше за удобство.
3. Исследования показывают, что наличие транспортной инфраструктуры, такой как метро, влияет на стоимость жилья. Это связано с тем, что такие объекты становятся более желанными для проживания.

Задание 1.4. Придумайте хотя бы 3 контрольные переменные, по крайней мере одна из которых должна быть бинарной и хотя бы одна – непрерывной. Кратко обоснуйте выбор каждой из них.

Ответ:
Нами были выбраны следующие непрерывные контрольные переменные:
1. Количество этажей в доме (floor)
Количество этажей в доме может влиять на стоимость жилья. В высотных зданиях квартиры на верхних этажах могут иметь лучшие виды и меньше шума, что делает их более привлекательными для покупателей. В то же время, в некоторых случаях квартиры на нижних этажах могут быть менее желательными из-за потенциального риска затопления или недостатка солнечного света.
2. Год постройки дома (year_built)
Год постройки влияет на состояние здания. Новостройки часто имеют современные удобства и технологии, что может повышать их стоимость по сравнению со старыми домами, требующими ремонта. Также новостройки могут находиться ближе к современным транспортным узлам.
3. Общая площадь квартиры (flat_area)
Общая площадь квартиры является одним из основных факторов, определяющих её стоимость. Большие квартиры обычно стоят дороже, так как предлагают больше пространства для жизни. Это также может влиять на спрос, так как семьи с детьми могут предпочитать более просторные квартиры.
4. Площадь кухни (kitchen)
Площадь кухни также важна для оценки стоимости жилья, особенно для семейных покупателей. Кухня является центральным местом в доме, и её размер может значительно влиять на комфорт проживания и общую привлекательность квартиры. Тем более сейчас люди часто предпочитают просторную кухню, совмещенную с гостиной.
5. Площадь ванной комнаты (bathroom)
Размер ванной комнаты может влиять на стоимость квартиры, так как это важное пространство для комфорта жильцов. Большие и хорошо оборудованные ванные комнаты могут повышать стоимость жилья.
6. Количество комнат (rooms)
Количество комнат в квартире напрямую связано с её функциональностью, возможностью размещения и комфортом. Большее количество комнат может увеличивать стоимость жилья, так как это позволяет удовлетворить потребности больших семей.
7. Этаж квартиры (floor_flat)
Этаж, на котором расположена квартира тоже играет немаловажную роль в определении стоимости квартиры, поскольку от этажа квартиры зависит вид из окна, уровень шума и доступность лифта. Квартиры на верхних этажах часто имеют лучшие виды, но могут быть менее удобными для людей с ограниченными возможностями, а иногда это просто вопрос предпочтений.

А в качестве бинарных контрольных переменных мы взяли:

1. Наличие ванной (bath)

Наличие ванной комнаты является бинарной переменной, которая может существенно влиять на стоимость жилья. Квартиры с ванной обычно считаются более желательными, чем те, которые имеют только душевую.

2. Тип ремонта (renovation)

Тип ремонта влияет на восприятие стоимости жилья и его привлекательность для покупателей. Квартиры с современным, дизайнерским ремонтом могут стоить значительно больше по сравнению с теми, которые имеют стандартный ремонт или только чистовую отделку.

3. Парковка (parking)

Наличие парковки является важным фактором для многих покупателей недвижимости, особенно в городах с высокой плотностью населения. Наличие подземной парковки в доме может повысить стоимость жилья.

4. Первичная продажа жилой недвижимости / Вторичная продажа (primary)

Первичные квартиры часто имеют более высокую цену из-за новизны и современных удобств.

5. Расположение жилья в пределах МКАДа (mkad) Расположение относительно МКАД влияет на доступность транспортных маршрутов и инфраструктуры. Квартиры внутри МКАД обычно имеют более высокую стоимость из-за лучшей транспортной доступности, развитой инфраструктуры по сравнению с районами за пределами кольца и меньших временных затрат на дорогу до центра.

Задание 1.5. Придумайте бинарную инструментальную переменную и обоснуйте, почему она удовлетворяет необходимым условиям.

Ответ:

В качестве бинарной инструментальной переменной мы выбрали наличие остановки общественного транспорта (автобус/трамвай) в шаговой доступности (10 мин ходьбы).

Наличие остановки общественного транспорта в пешей доступности предполагает, что жилье становится более доступным для потенциальных покупателей, что может привести к увеличению спроса и, соответственно, повышению цен на недвижимость.

Исследования показывают, что близость к остановкам общественного транспорта положительно влияет на стоимость жилья, так как это упрощает доступ к основным транспортным маршрутам и центру города.

Данная бинарная переменная может быть независимой от других факторов, влияющих на цену жилья. Например, она не зависит от индивидуальных характеристик квартиры (таких как площадь или количество комнат), что делает ее хорошим выбором в качестве инструментальной переменной. Это позволяет избежать смещения оценок.

Наличие остановки общественного транспорта непосредственно влияет на стоимость жилья. Исследования показывают, что квартиры, находящиеся вблизи остановок, могут стоить на 10-30% больше по сравнению с аналогичными объектами, расположенными дальше от общественного транспорта. Это создает достаточную силу связи между инструментальной переменной и зависимой переменной.

2 Генерация и предварительная обработка данных

Задание 2.1. Опишите математически предполагаемый вами процесс генерации данных. Примечание: оценивается в том числе оригинальность предложенного вами процесса, поэтому, в частности, не рекомендуется использовать совсем простые линейные модели.

Ответ:
Предположим, что условная вероятность факта наличия автобусной или трамвайной остановки около жилого дома положительно связана с годом постройки здания, первичной/вторичной продажей жилья, расположения квартиры в пределах МКАДа и наличия парковки:

$$\begin{gathered}\mathrm{P}\left(\mathrm{bus}_i = 1 | \mathrm{year\_built}_i, \mathrm{primary}_i, \mathrm{mkad}_i, \mathrm{parking}_i\right) = \\ = \Phi\left(\underbrace{\frac{0.31 * \mathrm{year\_built}_i}{(120 - 1.12 * \mathrm{primary}_i - \mathrm{mkad}_i)} + 0.85 * \mathrm{mkad}_i \times \mathrm{priming}_i - 4.5}_{\text{индекс}}\right)\end{gathered}$$

Где $\Phi()$ - функция распределения стандартного нормального распределения.

Удобно предположить, что условные вероятности переменной воздействия зависят от контрольных переменных, инструментальной переменной и ненаблюдаемой переменной.

$$\begin{gathered}P(\mathrm{subway}_i = 1 | \mathrm{year\_built}_i, \mathrm{mkad}_i, \mathrm{primary}_i, \mathrm{quality}_i, \mathrm{bus}_i) = \\ = F_{\text{Logistic}}\left(2.8 \times \ln\left(\mathrm{year\_built}_i + 0.01\right) + \sqrt{\mathrm{bus}_i} + 0.54 * \mathrm{mkad}_i \times \mathrm{primary}_i + 0.93 \times \mathrm{quality}_i - 77\right)\end{gathered}$$

Где F_{Logistic} - функция распределения стандартного логистического распределения.
Для краткости введем обозначение для условной вероятности наличия станции метро при наличии/отсутствии остановки общественного транспорта:

$$p_k^{\mathrm{subway}_i} = P(\mathrm{subway}_i = 1 | \mathrm{year\_built}_i, \mathrm{mkad}_i, \mathrm{primary}_i, \mathrm{quality}_i, \mathrm{subway}_i = k), \text{ где } k \in \{0, 1\}$$

Для того, чтобы впоследствии анализировать локальные средние эффекты воздействия LATE, необходимо различать величину переменной воздействия subway_i в зависимости от значения инструмента bus_i. Для этого рассмотрим ни от чего не зависящую равномерную случайную величину $U_i \sim U(0, 1)$ и введем гипотетические переменные:

$$\mathrm{subway}_{1i} = I(p_1^{\mathrm{bus}_i} \geq U_i)$$

$$\mathrm{subway}_{0i} = I(p_0^{\mathrm{bus}_i} \geq U_i)$$

$$I(\text{условие}) = \begin{cases} 1, \text{ если условие выполнено} \\ 0, \text{ в противном случае} \end{cases}$$

Переменные subway_{1i} и subway_{0i} отражают потенциальную возможность наличия метро около квартиры при наличии остановки общественного транспорта.

Для того, чтобы впоследствии анализировать локальные средние эффекты воздействия LATE, необходимо различать величину переменной воздействия subway_i в зависимости от значения инструмента bus_i.
Для этого рассмотрим ни от чего не зависящую равномерную случайную величину $U_i \sim U(0,1)$ и введем гипотетические переменные:

$$\text{subway}_{1i} = I(P(\text{subway}_i = 1 | \text{other X}, \text{bus}_i = 1) \geq U_i)$$
$$\text{subway}_{0i} = I(P(\text{subway}_i = 1 | \text{other X}, \text{bus}_i = 0) \geq U_i)$$

Где:

$$I(\text{условие}) = \begin{cases} 1, \text{ если условие выполнено} \\ 0, \text{ в противном случае} \end{cases}$$

Наблюдаемое наличие станции метро можно выразить как:

$$\text{subway}_i = \begin{cases} \text{subway}_{1i}, \text{ если } \text{bus}_i = 1 \\ \text{subway}_{0i}, \text{ если } \text{bus}_i = 0 \end{cases}$$

Напомним, что к соблюдателям относятся те, у кого $\text{subway}_{1i} > \text{subway}_{0i}$, то есть наличие станции метро при $\text{bus}_i = 1$ и ее отсутствие - при $\text{bus}_i = 0$.
Таким образом, квартиры можно разделить на 4 группы:
Always takers - те, у кого $\text{subway}_{0i} = \text{subway}_{1i} = 1$: строят станцию метро независимо от наличия/отсутствия остановки общественного траснпорта.
Never takers - те, у кого $\text{subway}_{0i} = \text{subway}_{1i} = 0$: не строят станцию метро независимо от наличия/отсутствия остановки общественного транспорта.
Compliers - те, у кого $\text{subway}_{1i} = 1$ и $\text{subway}_{0i} = 0$, то есть $\text{subway}_{1i} > \text{subway}_{0i}$: строят станцию метро лишь в случае, если есть остановка общественного транспорта.
Deniers- те, у кого $\text{subway}_{1i} = 0$ и $\text{subway}_{0i} = 1$, то есть $\text{subway}_{1i} < \text{subway}_{0i}$: строят станцию метро лишь в случае, если нет остановки общественного транспорта.
Для соблюдения предпосылок используемых методов важно отсутствие Deniers, что гарантируется используемым процессом генерации данных.
Наблюдаемый в данных факт наличия метро можно выразить как:

$$\text{subway}_i = \begin{cases} \text{subway}_{1i}, \text{ если } \text{bus}_i = 1 \\ \text{subway}_{0i}, \text{ если } \text{bus}_i = 0 \end{cases} \quad = \; = \text{subway}_{1i} \times \text{bus}_i + \text{subway}_{0i} \times (1 - \text{bus}_i)$$

При слабой корреляции между subway_i и quality_i проблема эндогенности окажется несущественной, а при слишком большой скорректировать эндоенность окажется чрезвычайно сложно.
При слабой корреляции между subway_i и bus_i инструмент не будет релевантным и поэтому не позволит скорректировать эндогенность.
Таким образом, сделаем так, чтобы корреляции находилась в некотором разумном диапазоне:

$$0.8 \geq |\text{Corr}\left(\text{subway}_i, \text{quality}_i\right)| \geq 0.2$$

$$0.8 \geq |\text{Corr}\left(\text{subway}_i, \text{bus}_i\right)| \geq 0.2$$

Задание 2.2. Симулируйте данные в соответствии с предполагаемым вами процессом и приведите корреляционную матрицу, а также таблицу со следующими описательными статистиками:
• Для непрерывных переменных: выборочное среднее, выборочное стандартное отклонение, медиана, минимум и максимум.
• Для бинарных переменных: доля и количество единиц.
Указания:
• Необходимо сгенерировать не менее 1000 наблюдений.
• Доля единиц не должна быть меньше 0.1 или больше 0.9 ни для одной из бинарных переменных.

Ответ:

Таблица 1: Описательные статистики переменных

	mean	std	min	median	max
price	371192	148876	13875	369713	1033918
subway	1	0	0	1	1
bus	1	0	0	1	1
floor	25	10	1	25	66
year_built	1980	10	1945	1980	2021
flat_area	80	94	15	47	500
kitchen	20	5	8	20	39
bathroom	5	1	2	5	9
rooms	3	2	1	3	10
floor_flat	25	10	1	25	61
bath	1	0	0	1	1
renovation	1	1	0	1	3
parking	1	1	0	1	2
primary	0	0	0	0	1
mkad	1	0	0	1	1

Таблица 2: Описательные статистики бинарных переменных

	Доля	Количество единиц
P(subway = 1)	0.5593	5593
P(bus = 1)	0.7734	7734
P(primary = 1)	0.3007	3007
P(mkad = 1)	0.7002	7002
P(bath = 1)	0.6031	6031

Таблица 3: Корреляционная матрица переменных

	subway	bus	floor	year_built	flat_area
subway	1.00	0.57	0.01	0.01	0.02
bus	0.57	1.00	0.00	0.03	0.00
floor	0.01	0.00	1.00	0.01	-0.00
year_built	0.01	0.03	0.01	1.00	-0.03
flat_area	0.02	0.00	-0.00	-0.03	1.00
kitchen	-0.01	0.00	-0.01	-0.00	0.00
bathroom	-0.00	0.01	0.01	0.00	-0.02
rooms	0.02	0.01	-0.00	-0.01	-0.01
floor_flat	0.01	-0.01	0.00	0.01	-0.01
primary	0.08	0.12	-0.00	-0.00	0.01
mkad	0.02	0.04	-0.00	0.00	0.01

	kitchen	bathroom	rooms	floor_flat	primary	mkad
subway	-0.01	-0.00	0.02	0.01	0.08	0.02
bus	0.00	0.01	0.01	-0.01	0.12	0.04
floor	-0.01	0.01	-0.00	0.00	-0.00	-0.00
year_built	-0.00	0.00	-0.01	0.01	-0.00	0.00
flat_area	0.00	-0.02	-0.01	-0.01	0.01	0.01
kitchen	1.00	0.00	0.01	0.00	0.01	-0.01
bathroom	0.00	1.00	-0.00	-0.01	0.00	-0.01
rooms	0.01	-0.00	1.00	0.00	0.01	-0.00
floor_flat	0.00	-0.01	0.00	1.00	-0.01	0.02
primary	0.01	0.00	0.01	-0.01	1.00	-0.00
mkad	-0.01	-0.01	-0.00	0.02	-0.00	1.00

Задание 2.3. Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20% до 30% наблюдений.

Ответ: Для дальнейшего анализа мы разделили нашу выборку на обучающую и тестовую в соотношении 80/20.

3 Классификация

Для задачи классификации мы использовали следующие методы: метод k-ближайших соседей, наивный Байесовский классификатор, деревья решений, случайный лес, логистическая регрессия, градиентный бустинг.

Задание 3.1. Отберите признаки, которые могут быть полезны при прогнозировании переменной воздействия и обоснуйте выбор каждой из них. Не включайте в число этих признаков целевую переменную.

Ответ: Для прогнозирования переменной воздействия (наличие метро в шаговой доступности) мы использовали следующие признаки:

- наличие остановки общественного транспорта в шаговой доступности (bus)
- количество этажей (floor)
- год постройки дома (year_built)
- общая площадь квартиры (flat_area)
- площадь кухни (kitchen)
- площадь ванной комнаты (bathroom)
- количество комнат (rooms)
- этаж квартиры (floor_flat)
- наличие ванной комнаты (bath)
- тип ремонта (renovation)
- парковка (parking)
- первичная/вторичная продажа жилой недвижимости (primary)
- расположение жилья в пределах МКАДа (mkad)

Задание 3.2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов: на обучающей выборке, на тестовой выборке, с помощью кросс-валидации (используйте только обучающую выборку). Проинтерпретируйте полученные результаты.
Ответ:

Таблица 4: Сводная таблица методов классификации

	ACC-train	ACC-test	ACC-CV-train
k-ближайших соседей	0.76900	0.76100	0.76475
Наивный Байесовский классификатор	0.76763	0.76700	0.76763
Деревья решений	0.76763	0.76700	0.76738
Случайный лес	0.76800	0.76700	0.76750
Логистическая регрессия	0.76763	0.76700	0.76763
Градиентный бустинг	0.77225	0.76400	0.76100

Сравнивая точность прогнозов на обучающей и тестовой выборках, представленные в Таблице 4, можно заметить, что в нашем случае не наблюдается переобучения, поскольку accuracy на тестовой выборке всегда меньше accuracy на обучающей для всех моделей классификации. Также и с кросс-валидацией на обучающей выборке, переобучения на наблюдается.

Сравнивая точность прогнозов на тестовой выборке для разных моделей, можем сделать вывод, что лучшей являются 4 модели: наивный Байесовский классификатор, деревья решений, случайный лес, логистическая регрессия.

Задание 3.3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте точность ACC. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны: изначальные и подобранные значения гиперпараметров, кросс-валидационная точность на обучающей выборке с исходными и подобранными значениями гиперпараметров, точность на тестовой выборке с исходными и подобранными значениями гиперпараметров. Проинтерпретируйте полученные результаты и далее используйте методы с подобранными значениями гиперпараметров.

Ответ: Для каждого метода классификации мы осуществили подбор гиперпараметров с помощью поиска по сетке (GridSearch). Результаты точностей моделей представлены в Таблицах 5,6.
Исходя из полученных метрик качества (accuracy) на тестовой выборке моделей с изначальными и подобранными гиперпараметрами, мы можем сделать следующие выводы:

- Нам удалось улучшить модель k-ближайших соседей, случайный лес и градиентный бустинг
- Что касается остальных моделей (наивный Байесовский классификатор, деревья решений, логистическая регрессия), подбор гиперпараметров не превзошел изначальные значения параметров.
- Сравнивая кросс-валидационную точность на обучающей выборке с точностью на тестовой выборке, переобучения моделей не наблюдается.

Повышенная сложность: Подберите на обучающей выборке оптимальные значения гиперпараметров случайного леса ориентируясь на значение OOB (out-of-bag) ошибки. Сопоставьте гиперпараметры и точность на тестовой выборке для случайного леса в зависимости от того, используется кросс-валидация или OOB ошибка. Объясните преимущества и недостатки OOB ошибки по сравнению с кросс-валидацей.

Ответ: Мы подобрали оптимальные значения гиперпараметров модели случайного леса на основе ошибки неотобранных элементов. Результаты представлены в Таблице 7.
Таким образом, можем заметить, что на основе ошибки неотобранных элементов подобрались такие же гиперарараметры, как и на кросс-валидации. При этом значения accuracy достаточно близки между собой.
Преимущество OOB ошибки - обычно быстрее, чем кросс-валидация, поскольку не нужно многократно обучать модель.
Недостаток OOB ошибки - нельзя использовать за пределами ансамблевых методов, что осложняет сопоставление их результатов с неансамблевыми методами.

Таблица 5: Сводная таблица методов классификации

	изначальные параметры	подобранные параметры
k-ближайших соседей	leaf_size: 30 metric: minkowski n_neighbors: 30 p: 2	n_neighbors: 32 p: 2
Наивный байесовский классификатор	var_smoothing: $1e-09$	var_smoothing: $1.8738e-05$
Деревья решений	criterion: entropy max_depth: 3 min_samples_leaf: 1 min_samples_split: 2	max_depth: 1
Случайный лес	bootstrap: True criterion: entropy max_depth: 12 max_features: sqrt max_samples: 500 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 100 oob_score: True random_state: 777	max_depth: 5 max_features: 9 min_samples_leaf: 1 min_samples_split: 3 n_estimators: 10
Логистическая регрессия	C: 1.0 fit_intercept: True intercept_scaling: 1 max_iter: 100 penalty: l2 solver: lbfgs	C: 0.000263665 max_iter: 500
Градиентный бустинг	criterion: friedman_mse learning_rate: 0.5 loss: log_loss max_depth: 2 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 50 random_state: 123 subsample: 1.0	learning_rate: 0.1 max_depth: 2 min_samples_leaf: 3 min_samples_split: 2 subsample: 0.75

Таблица 6: Сводная таблица методов классификации

	ACC-CV-initial	ACC-CV-GS	ACC-test-initial	ACC-test-GS
k-ближайших соседей	0.76475	0.76562	0.76100	0.76200
Наивный Байесовский классификатор	0.76763	0.76763	0.76700	0.76700
Деревья решений	0.76738	0.76763	0.76700	0.76700
Случайный лес	0.76750	0.76838	0.76700	0.76750
Логистическая регрессия	0.76763	0.76763	0.76700	0.76700
Градиентный бустинг	0.76100	0.76763	0.76400	0.76700

Таблица 7: Сравнение OOB и CV для подбора параметров

	OOB	CV
подобранные гиперпараметры	max_depth: 5 max_features: 9 min_samples_leaf: 1 min_samples_split: 3 n_estimators: 10	max_depth: 5 max_features: 9 min_samples_leaf: 1 min_samples_split: 3 n_estimators: 10
ACC-test	0.76838	0.76750

Задание 3.4. Постройте ROC-кривую для ваших моделей и сравните их по AUC на тестовой выборке.

Ответ: Для всех моделей были построены графики ROC-кривых и значения AUC. Графики представлены ниже (Рис.1-5).

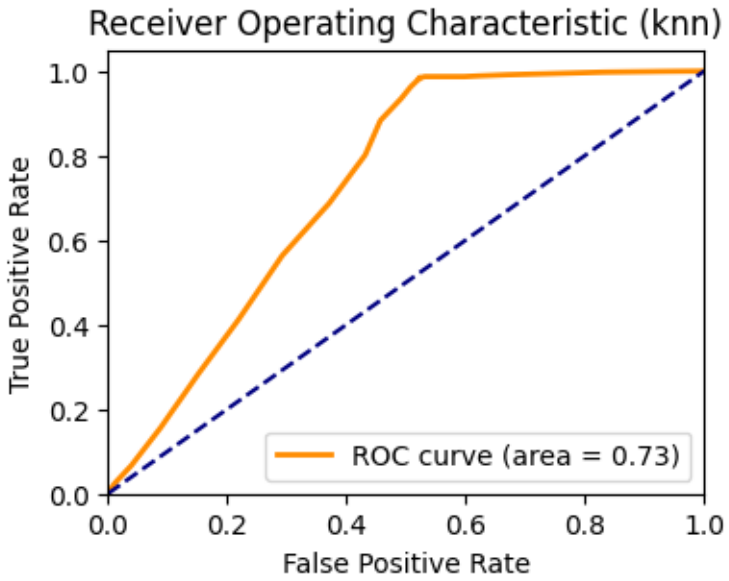

Рис.1. ROC-кривая и значение AUC для модели k-ближайших соседей

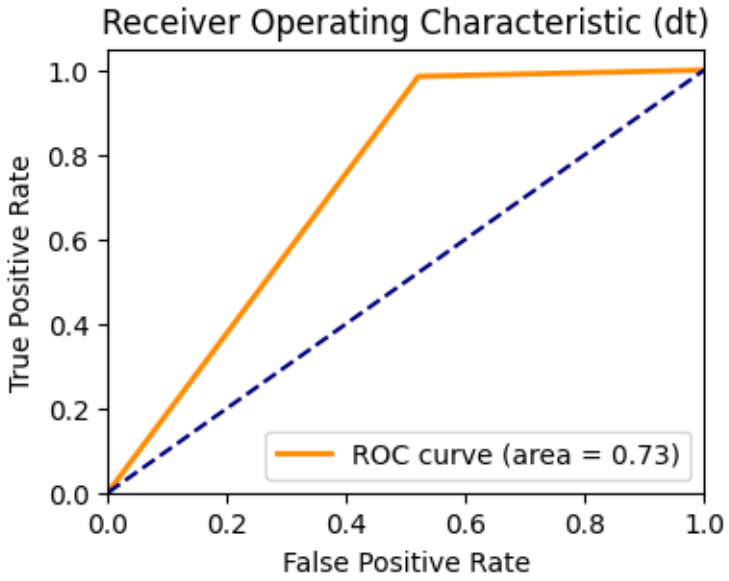

Рис.2. ROC-кривая и значение AUC для модели решающих деревьев

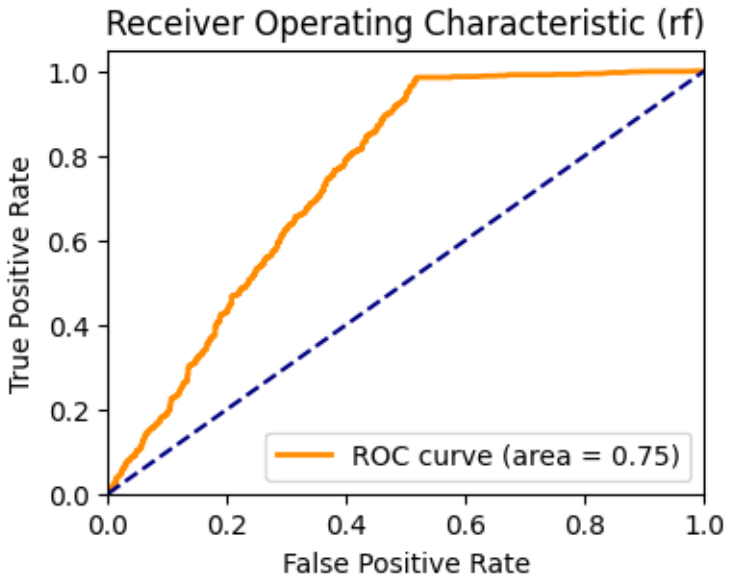

Рис.3. ROC-кривая и значение AUC для модели случайного леса

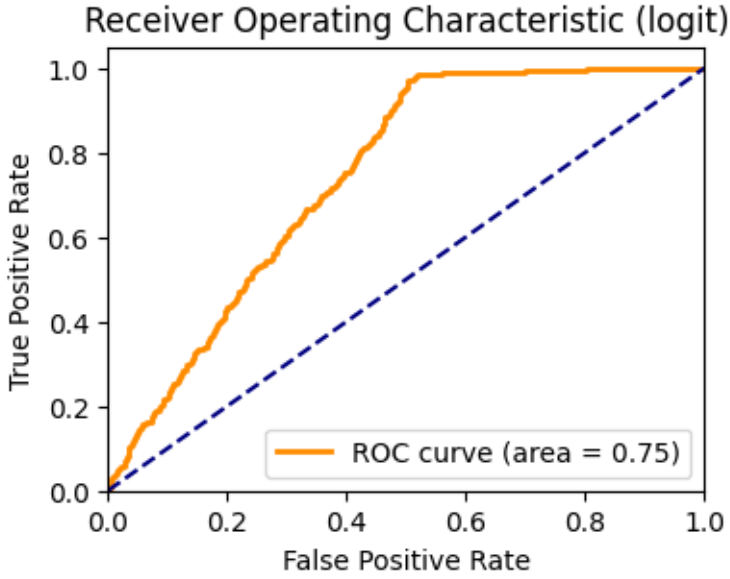

Рис.4. ROC-кривая и значение AUC для модели логистической регрессии

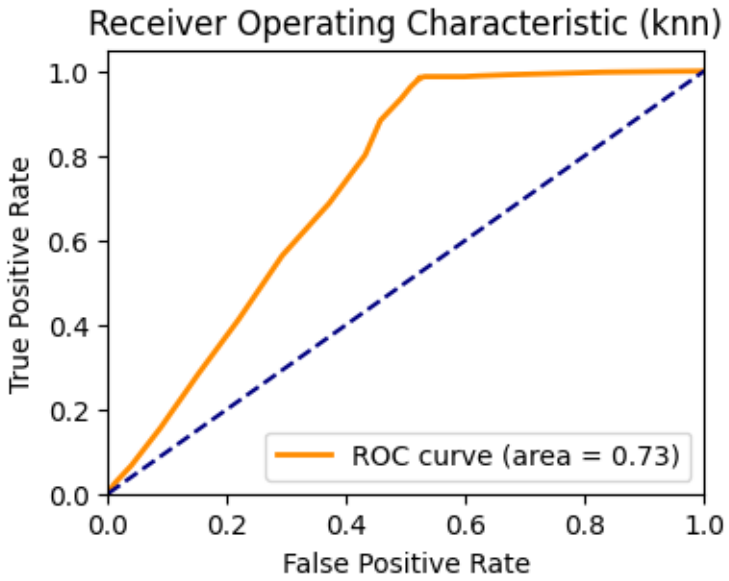

Рис.5. ROC-кривая и значение AUC для модели градиентного бустинга

Можно заметить, что значения площади под ROC-кривой для всех используемых моделе достаточно близки. Однако лучшими являются случайный лес и логистическая регрессия (AUC=0.75).

Задание 3.5. Постройте матрицу ошибок и предположите цены различных видов прогнозов. Исходя из критерия максимизации прибыли на обучающей выборке подберите оптимальный порог прогнозирования для каждого из методов и сравните прибыли на тестовой выборке при соответствующих порогах. Результат представьте в форме таблицы, в которой должны быть указаны как AUC, так и прибыли (на тестовой выборке). Проинтерпретируйте полученный результат.

Ответ: Матрицы ошибок для всех моделей представлены ниже (Рис.6-10).

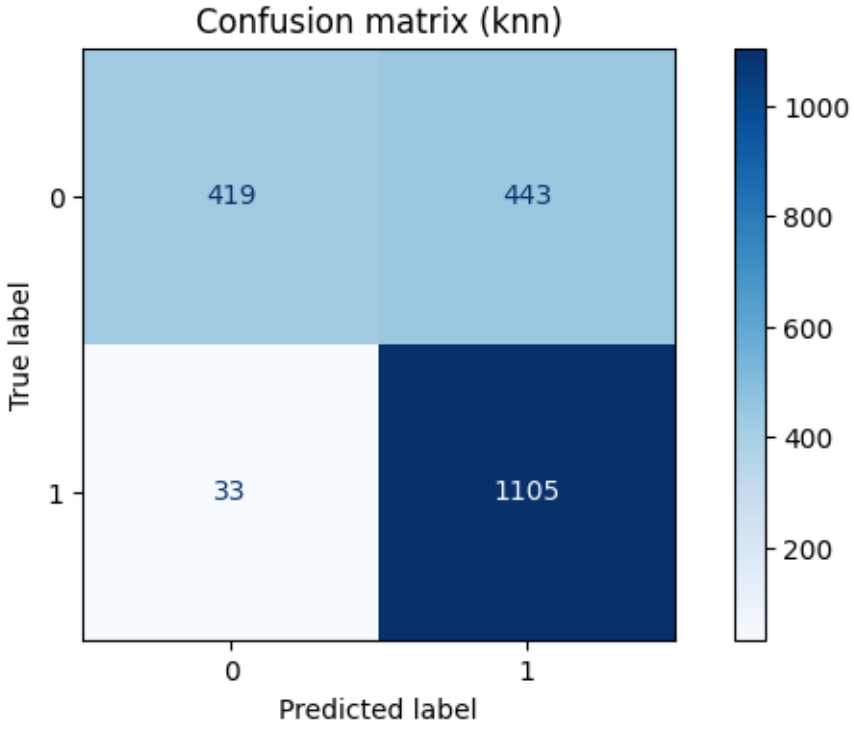

Рис.6. Матрица ошибок для модели k-ближайших соседей

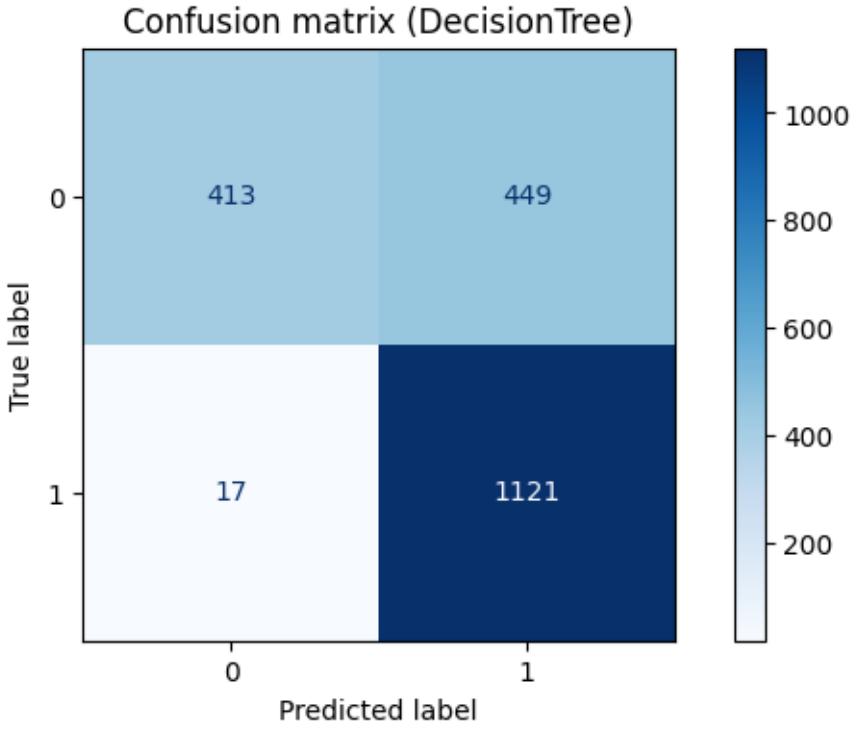

Рис.7. Матрица ошибок для модели решающих деревьев

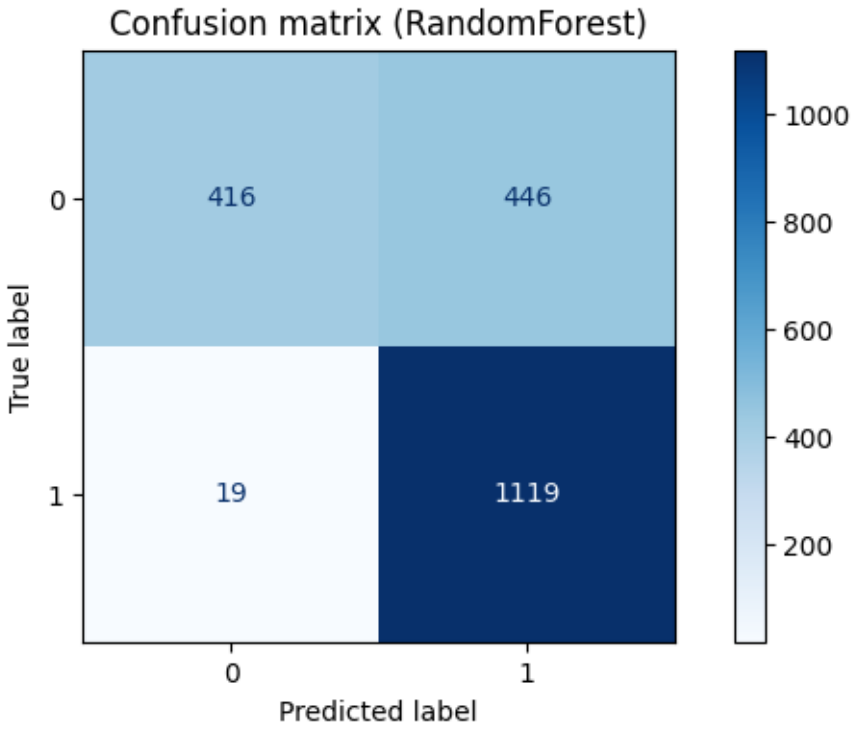

Рис.8. Матрица ошибок для модели случайного леса

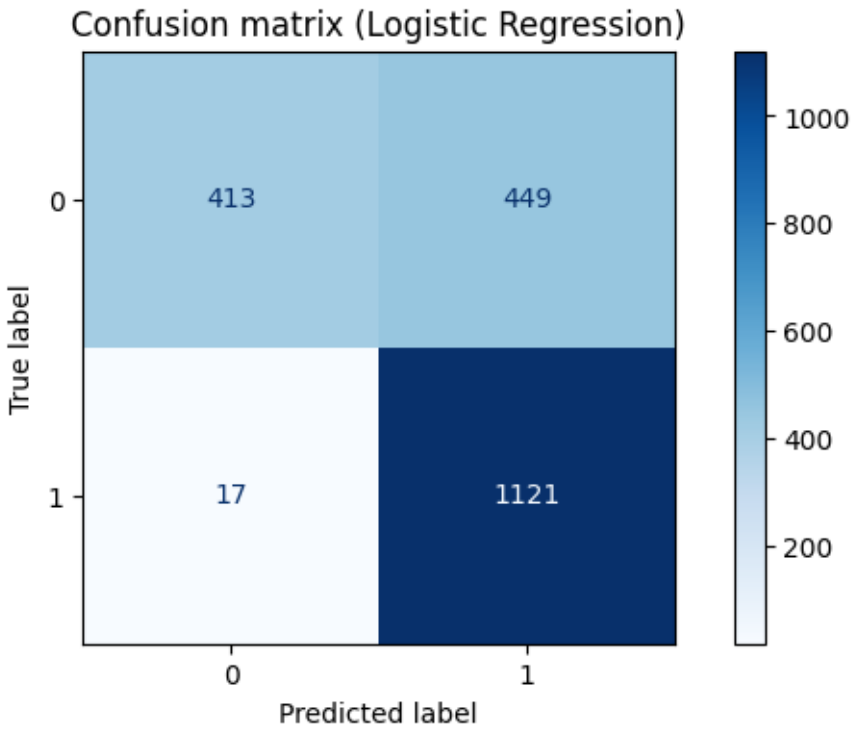

Рис.9. Матрица ошибок для модели логистической регрессии

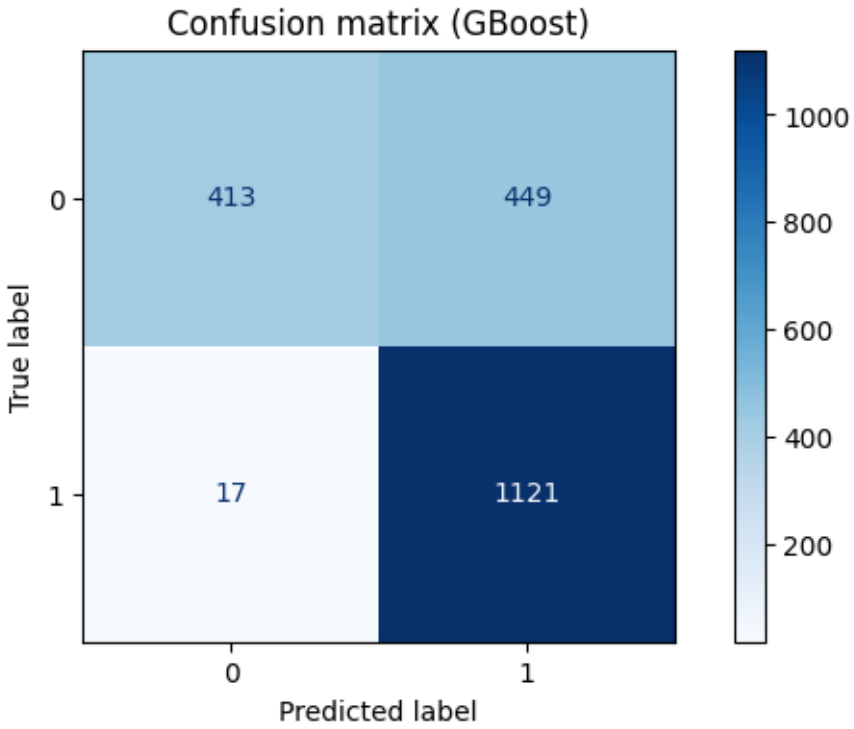

Рис.10. Матрица ошибок для модели градиентного бустинга

Можно заметить, что сумма ложноотрицательных (FN) и ложноположительных (FP) ошибок схожа в разных моделях. Однако лучшей является случайный лес, поскольку показывает наименьшую сумму FN+FP=465.

Задание 3.6 На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор.

Ответ: В результате проведенного анализа различных методов классификации для решения задачи предсказания, модель случайного леса была определена как наилучший классификатор.
Модель случайного леса продемонстрировала высокую точность на тестовой выборке, что свидетельствует о ее способности правильно классифицировать значительное количество наблюдений. Высокий уровень точности указывает на то, что модель эффективно справляется с задачей различения классов, минимизируя количество ошибок.
Значение ROC AUC для модели случайного леса также было высоким, что подтверждает ее способность различать положительные и отрицательные классы. Площадь под кривой ROC является важной метрикой, так как она учитывает как истинные положительные, так и ложные положительные результаты, предоставляя более полное представление о качестве модели.
Анализ матрицы ошибок показал, что модель случайного леса имеет низкий уровень ложных определений и пропусков. Это означает, что модель не только точно классифицирует большинство наблюдений, но и минимизирует количество ошибок в критически важных классах, что особенно важно в задачах с несбалансированными классами.

Тем не менее, в контексте данного анализа худшим классификатором оказался метод k-ближайших соседей (k-NN).
Метод k-ближайших соседей показал чуть более низкую точность по сравнению с другими моделями. Это может быть связано с тем, что k-NN чувствителен к шуму в данных и может неправильно классифицировать наблюдения из-за близости к неправильно размеченным объектам.

4 Регрессия

Задание 4.1. Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия. Содержательно обоснуйте выбор признаков.

Ответ: Отберем признаки, которые могут быть полезны при прогнозировании цены на квадратный метр.
Целевая переменная:

- Цена 1 кв м жилой недвижимости (price)

Мы выбрали эти переменные, так как логично предполагать, что характеристики квартиры будут влиять на ее стоимость:
Контрольные переменные:

- Количество этажей в доме (floor)
- Год постройки дома (year_built)
- Общая площадь квартиры (flat_area)
- Площадь кухни (kitchen)
- Площадь ванной комнаты (bathroom)
- Количество комнат (rooms)
- Этаж квартиры (floor_flat)
- Наличие ванной (bath)
- Тип ремонта (renovation)
- Парковка (parking)
- Первичная продажа жилой недвижимости / Вторичная продажа (primary)
- Расположение жилья в пределах МКАДа (mkad)

Ненаблюдаемые переменные, порождающие эндогенность:

- Качество инфраструктуры района (quality):

Качество инфраструктуры района включает в себя общее состояние дорог, наличие парков, образовательных учреждений и других социальных услуг. Высокое качество инфраструктуры может способствовать повышению цен на жилье. Так как качество инфраструктуры не наблюдается напрямую, оно может создать смещение в оценках.

Задание 4.2. Для каждого метода с помощью кросс-валидации на обучающей выборке под- берите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте RMSE. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

- изначальные и подобранные значения гиперпараметров.
- кросс-валидационное значение RMSE на обучающей выборке с исходными и подобранными значениями гиперпараметров.
- значение RMSE на тестовой выборке с исходными и подобранными значениями гиперпараметров.

Проинтерпретируйте полученные результаты.

Ответ: Наилучшие результаты показала регрессия с градиентным бустингом. После подбора параметров для модели градиентного бустинга RMSE и MAPE на тестовой выборке стали немного ниже, чем на тренировочной, что свидетельствует об отсутствии переобучения.

Таблица 8: Изначальные и подобранные параметры

	изначальные параметры	подобранные параметры
МНК	'fit_intercept': True 'positive': False	'fit_intercept': True 'positive': True
Случайный лес	'bootstrap': True 'criterion': 'squared_error' 'max_depth': 2 'max_features': 1.0 'min_samples_leaf': 1 'min_samples_split': 2 'n_estimators': 100	'max_depth': 15 'max_features': 9 'n_estimators': 400
k ближайших соседей	'leaf_size': 30 'metric': 'minkowski' 'n_neighbors': 8 'p': 2 'weights': 'distance'	'n_neighbors': 8 'p': 2 'weights': 'distance'
Градиентный бустинг	'alpha': 0.9 'criterion': 'friedman_mse' 'learning_rate': 0.1 'loss': 'squared_error' 'max_depth': 3 'min_samples_leaf': 1 'min_samples_split': 2 'n_estimators': 100	'learning_rate': 0.1 'max_depth': 4 'min_samples_leaf': 2 'min_samples_split': 6 'subsample': 0.75

Таблица 9: Метрики RMSE регрессии

	RMSE на train	RMSE на test	RMSE на train CV
МНК	826320.33	13072.79	12850.40
Случайный лес	51696.38	51970.06	51272.50
k ближайших соседей	0.00	60976.80	64128.86
Градиентный бустинг	9484.27	10922.65	10614.68

Таблица 10: Метрики MAPE регрессии

	MAPE на train	MAPE на test	MAPE на train CV
МНК	2.83	0.03	0.04
Случайный лес	0.23	0.20	0.23
k ближайших соседей	0.00	0.24	0.29
Градиентный бустинг	0.02	0.02	0.02

На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор Самой плохой результат показали модели k-ближайших соседей и случайный лес. Так как после подбора параметров ошибка RMSE и MAPE остались самыми большими на фоне моделей GBoost и МНК, а также присутствуют признаки переобучения: на тренировочной выборке значения метрик ниже, чем на тестовой (чем ниже, тем лучше предсказывает модель).

5 Эффекты воздействия

Для выполнения данного задания мы объединили обучающую и тестовую выборки в одну.
Задание 5.1. Математически запишите и содержательно проинтерпретируйте потенциальные исходы целевой переменной. Объясните, как они связаны с наблюдаемыми значениями целевой переменной.

Ответ:

$$\text{price}_i = \begin{cases} \text{price}_{1i}, \text{ если } \text{subway}_i = 1 \\ \text{price}_{0i}, \text{ если } \text{subway}_i = 0 \end{cases} =$$

$$= \text{price}_{1i} \times \text{subway}_i + \text{price}_{0i} \times (1 - \text{subway}_i)$$

$$\text{price}_i = \alpha + \text{subway}_i + g(X_i)$$

Мы ожидаем, что цена за квадратный метр на квартиру, около которой есть метро будет выше, чем цена за квадратный метр на эту же квартиру, если бы станции метро рядом не было. Исходя из данных, мы наблюдаем только одно из состояний квартиры: либо около нее уже есть станция метро, либо ее нет. Поэтомы мы будем оценивать цену на квадратный метр для квартиры, около которой есть метро, в ситуации, если бы станции метро не было и наоборот.

Задание 5.2. Используя симулированные вами, но недоступные в реальных данных потенциальные исходы (гипотетические значения), получите оценки среднего эффекта воздействия, условных средних эффектов воздействия и локального среднего эффекта воздействия. Для ATE и LATE результаты представьте в форме таблицы, а для CATE постройте гистограмму или ядерную оценку функции плотности. Проинтерпретируйте полученные значения.

Ответ:

Таблица 11: Истинные эффекты воздействия

	Истинное значение
ATE	22523.71
LATE	22632.80
ATET	22694.11

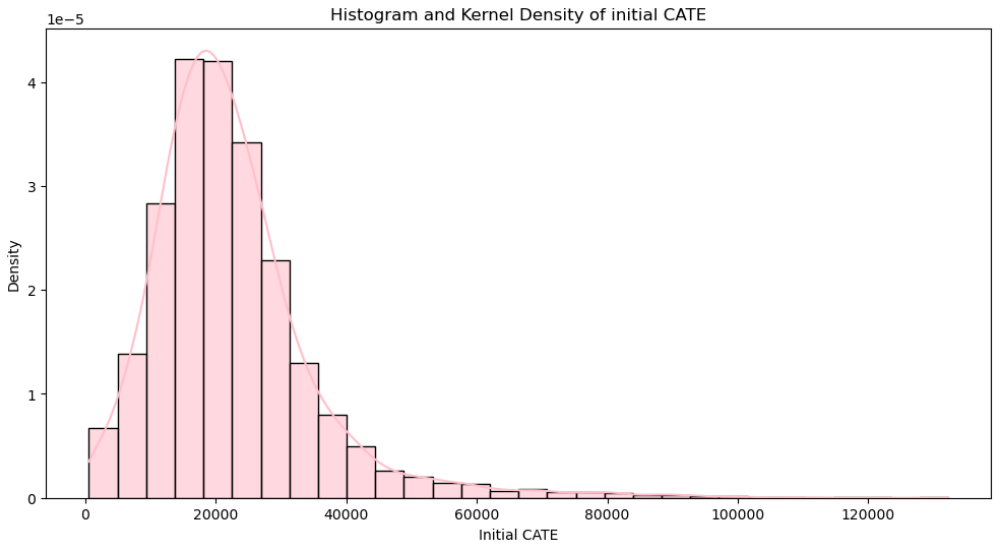

Рис.7. Гистограмма условных средних эффектов воздействия (CATE)

Средний эффект воздействия (ATE) показывает, что в среднем если около квартиры в шаговой доступности (10 мин) есть станция метро, то она будет стоить дороже на ATE =22523.71 руб., чем если бы у этой же квартиры рядом не было станции метро.

Локальный средний эффект воздействия (LATE) показывает, ATE среди соблюдателей (Compliers). Compliers - те, у кого subway_1 =1 и subway_0 = 0, то есть subway_1 > subway_0: строят станцию метро лишь в случае, когда есть остановка общественного транспорта. При отсутствии отрицателей, наличи соблюдателей и экзогенном инструменте Zi для оценивания LATE достаточно оценить E(Yi |Zi) и P(Ti |Zi).

То есть LATE показывает, что в среднем если около квартиры в шаговой доступности (10 мин) есть станция метро при наличие автобусной остановки в реальности, то она будет стоить дороже на LATE = 22632.80 руб., чем если бы у этой же квартиры рядом не было метро, так как нет автобусной остановки.

LATE = E(Y1i-Y0i | T1i > T0i)

Средний эффект воздействия на подвергшихся воздействию ATET показывает,что квартира рядом с метро будет стоить ATET = на 22694.11 руб. дороже, если бы рядом не было станции метро среди только тех квартир, около которых уже есть метро.

ATET = E(Y1i - Y0i | Ti=1) = E(Y1i | T=1) - E(Y0i | T=1)

Задание 5.3. Оцените средний эффект воздействия как разницу в средних по выборкам тех, кто получил и не получил воздействие. Опишите недостатки соответствующего подхода с учетом специфики

рассматриваемой вами экономической проблемы.

Ответ:

Таблица 12: Сравнение истинного эффекта и наивного прогноза

	Оценка
ATE	22523.709
ATE_naive	24905.187

Наивный подход предполагает, что ATE оценивается как среднее от разницы цены за квадратный метр квартир, рядом с которыми есть метро, и квартир, рядом с которыми метро нет. Главный недостаток данного метода - это допущение о независимости наблюдений:

$$E(Y\_1i \mid T\_i = 1) = E(Y\_1i)$$
$$E(Y\_0i \mid T\_i = 0) = E(Y\_0i)$$

В наших данных такое допущение не выполняется, поскольку контрольные переменные (year built, bus, mkad, primary) и ненаблюдаемая переменная, вызывающая эндогенность (quality) связаны с переменной воздействия subway и целевой переменной price.

Задание 5.4. Используя оценки, полученные лучшими из обученных ранее классификационных и регрессионных моделей, оцените средний эффект воздействия с помощью:

- метода наименьших квадратов.
- условных математических ожиданий.
- взвешивания на обратные вероятности (в случае возникновения ошибок убедитесь в отсутствии оценок вероятностей, равных 0 или 1 и при необходимости измените метод оценивания).
- метода, обладающего двойной устойчивостью.
- двойного машинного обучения.

Сравните результаты и назовите ключевую предпосылку этих методов. Содержательно обсудите причины, по которым она может соблюдаться или нарушаться в вашем случае. Приведите содержательную экономическую интерпретацию оценки среднего эффекта воздействия.

Ответ: Теперь мы ослабляем предпосылку о независимости и вводим предпосылку об условной независимости (conditional mean independence):

$$E(Y\_1i \mid X\_i, T\_i = 1) = E(Y\_1i \mid X\_i)$$
$$E(Y\_0i \mid X\_i, T\_i = 0) = E(Y\_0i \mid X\_i)$$

При прочих равных контрольных переменных Xi информация о наличие рядом с квартирой станции метро Ti не влияет на наши ожидания по поводу того, сколько бы квартира стоила в случаях, когда около нее есть Y1i и когда около нее нет Y0i станции метро. Допущение об условной независимости будет соблюдаться, когда все контрольные переменные Xi отражают все факторы, которые могут быть статистически связаны с метро (Ti), наблюдаемыми ценами и с не наблюдаемыми ценами в альтернативном сценарии (Yji).

МНК

Оценим ATE как среднюю разницу в оценках цены за квадратный метр, полученных с помощью МНК, отдельно для квартир, около которых есть метро, и нет соответственно.

Строим две регрессии с помощью МНК:

1-ая: обучается на наблюдениях, когда около квартиры нет станции метро и прогнозирует оценка цены для каждой квартиры во вселенной, если рядом с квартирой нет метро ($price_0$ - МНК оценка E($price_0$ | X) для всех квартир.

2-ая: обучается на наблюдениях, когда около квартиры есть станция метро и прогнозируется оценка цены для каждой квартиры во вселенной, если рядом с квартирой есть метро ($price_1$ - МНК оценка E($price_1$ | X) для всех квартир).

Условные математические ожидания Оценим с помощью двух методов:

- Single-learner
- Two-learner

Для T-learner мы использовали модель регрессии GBoost, с подобранными гиперпараметрами, так как она дала более точную оценку для ATE по сравнению с истинным ATE. Мы дважды обучили модель: сначала на наблюдениях по квартирам, рядом с которыми не было станции метро, потом на наблюдениях, где было метро. Потом перекрестно прогнозировали:

Таблица 13: Сравнение оценок ATE разных моделей

	Оценка
ATE	22523.709
ATE_naive	24905.187
ATE_ls	22565.756
ATE_T-learner	22525.166
ATE_S-leaner	21762.794
ATE_IPW	24010.920
ATE_DR	22504.647
ATE_dml_standard	22449.537

Наиболее точной оценкой оказалась оценка T-learner и метод двойной устойчивости.

Задание 5.5. Оцените локальный условный эффект воздействия с помощью:

- двойного машинного обучения без инструментальной переменной.
- двойного машинного обучения с инструментальной переменной.

Сопоставьте результаты и объясните, в чем в вашем случае будет заключаться различие между средним эффектом воздействия и локальным средним эффектом воздействия. Приведите содержательную экономическую интерпретацию оценки локального среднего эффекта воздействия.

Ответ: Мы оценили локальные эффект воздействия без инструментальной переменной (bus - наличие автобусной остановки рядом с квартирой в шаговой доступности) и с ней.
LATE без инструментальной переменной неточная из-за эндогенности. Добавив, инструментальную переменную, мы снизили влияние эндогенности.

Таблица 14: Локальные средние эффекты воздействия с инструментом и без

	Оценка
ATE	22523.709
LATE	22632.800
LATE_dml_standard2	22009.730
LATE_dml_iv	22107.819

В нашем случае разница между ATE и LATE заключается в том, что средний эффект воздействия отражает среднюю разницу в цене за квадратный метр для каждой квартиры в случае, когда около нее есть метро и когда его нет. Локальный средний эффект воздействия отражает среднюю разницу в цене за квадратный метр для тех квартир, около которых было бы метро только в случае, если рядом есть автобусная остановка.
Задание 5.6. Оцените условные средние эффекты воздействия с помощью:

- метода наименьших квадратов.
- S-learner.
- T-learner.
- метода трансформации классов.
- X-learner.

Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при реализации государственных программ.

Ответ: Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при

реализации государственных программ.

Оценки CATE через МНК, S-learner, T-learner выглядят адекватно, так как условный средний эффект воздействия - это положительная величина с медианой около 20 000, что соответствует гипотезе о том, что наличие метро около квартиры увеличивает цену на квадратный метр квартиры примерно на 20 000 руб. (ATE = 22 523 руб.)

Метод трансформации классов показал сомнительные результаты, так как медиана оценки CATE около нуля. То есть эта оценка соответствует гипотезе, что наличие метро незначительно влияет на цену квадратного метра квартиры. Это может быть вызвано методом оценки обратных взвешенных вероятностей.

X-learner показал хороший результат, так как он отражает CATE в диапазоне от 22508 до 22540 руб., что соответствует истинному значению ATE = 22523.71. X-learner хорошо работает на группах воздействия, которые включают в себя малое число наблюдений, что осложняет оценивание E(Yi|Xi,Ti = 1) по данным группы воздействия.

Адекватные оценки данного метода можно объяснить тем, что наша выборка состоит только из 10 000 наблюдений по Москве и Московской области. В реальности же квартир в Москве и области значительно больше, поэтому несмотря на то, что в нашей выборке наблюдений, в которых квартир с метро рядом больше половины (56%), это еще не отражает всю генеральную совокупность. Однако выбор 10 000 наблюдений был связан с ограниченностью вычислительных мощностей.

Задание 5.7. Выберите лучшую модель оценивания условных средних эффектов воздействия, используя:

- истинные значения условных средних эффектов воздействия.
- прогнозную точность моделей.
- псевдоисходы.

Проинтерпретируйте различия в результатах различных подходов.

Ответ: Опираясь на истинные значения условных средних эффектов воздействия, наилучшие результаты показали оценки моделей X-learner (ATE = 22 522), T-learner (ATE = 22 525) и метод двойной устойчивости (ATE = 22 504).

Таблица 15: Истинные значения условных средних эффектов воздействия

	Оценка
ATE	22523.709
ATE_naive	24905.187
ATE_ls	22565.756
ATE_T-learner	22525.166
ATE_S-leaner	21762.794
ATE_X-leaner	22522.367
ATE_IPW	24010.920
ATE_DR	22504.647
ATE_dml_standard	22449.537

Прогнозную точность моделей мы посчитали по RSME. Наименьшая ошибка истинного RMSE в оценках метода S-learner (RMSE = 3384) и T-learner (RMSE = 3405).

Таблица 16: Прогнозная точность моделей (истинная RMSE

	Оценка
LS	3420
T-learner	3405
S-learner	3384
CT	532251
X-learner	12838
DR	4949

Также мы оценивали сравнение CATE на основании псевдо исходов через RMSE. Наименьшее RMSE оказалось у метода трансформации классов. Напомним, что этот метод единственный оценивал CATE близкое к нулю, то есть в пользу гипотезы о том, что наличие метро не оказывает существенного влияния на цену квадратного метра квартиры. В то время как все остальные методы оценивали влияние метро около 22 000 руб., что близко к истинному значению среднего эффекта воздействия (ATE = 22 523) Следующие по минимальным RMSE после метода трансформации - это методы двойной устойчивости и T-learner, что немного согласуется с выводами по оценке ATE.

Таблица 17: Прогнозная точность моделей (оценка RMSE

	Оценка
LS	797055
T-learner	796867
S-learner	796997
CT	558609
X-learner	797208
DR	796540

Таким образом, мы отдаем предпочтение методам двойной устойчивости и T-learner. X-learner также показал хороший результат при оценке ATE и CATE, но хуже показал себя при оценке прогнозной точности моделей.