Национальный исследовательский университет

“Высшая школа экономики”

Факультет экономических наук

Направление: “Экономика”

Домашняя работа по курсу «Машинное обучение в экономике»

Влияние количества свободных мест на борту на цену авиабилетов

Работу выполнили:

Золотухина Евгения, группа 9,

семинарист: Столяров М.Э.

Рыбин Сергей, группа 4,

семинарист: Погорелова П.В.

Москва 2024

1 Обоснование темы

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

Целевая переменная - цена авиабилета (в $), Ticket_Price.

Переменная воздействия - количество свободных мест (бинарная, 1 в случае, если свободных мест меньше 20, 0 - иначе), Last_ticket.

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.

Изучение влияния количества свободных мест на цены авиабилетов может быть полезно для различных целей. В первую очередь данная информация полезна для **бизнеса**, а именно для авиакомпаний, результаты исследования могут помочь в разработке более эффективной ценовой стратегии. Понимание того, как изменение количества свободных мест влияет на цену билетов, позволяет компаниям корректировать цены для максимизации прибыли и заполненности рейсов. Это также помогает в управлении загрузкой, так как знание того, что билеты на рейсы с малым количеством свободных мест стоят дороже, позволяет лучше управлять ресурсами и предлагать скидки на менее популярные рейсы для стимулирования спроса. Кроме того, информация о ценовой эластичности в зависимости от свободных мест может быть использована для разработки маркетинговых кампаний и программ лояльности, направленных на увеличение продаж в периоды низкого спроса.

Безусловно, информация о влиянии количества свободных мест полезна и для **пассажиров**: они могут использовать понимание механизма ценообразования для лучшего планирования своих поездок и выбора оптимального времени для покупки билетов. Это позволяет пассажирам экономить деньги, бронируя билеты заранее или выбирая менее загруженные рейсы.

Государственные органы также могут извлечь пользу из результатов такого исследования. Они могут использовать данные для разработки рекомендаций по регулированию цен на авиабилеты, чтобы предотвратить чрезмерное завышение цен в условиях высокого спроса и малого количества свободных мест. Понимание динамики цен способствует защите прав пассажиров, обеспечивая прозрачность ценообразования и предупреждая необоснованные повышения цен в случае ограниченного предложения. Анализ спроса и цен на авиабилеты также может помочь в планировании развития аэропортов и транспортной инфраструктуры, чтобы лучше удовлетворять потребности пассажиров.

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

Исследование причинно-следственной связи между зависимой переменной (ценой авиабилета) и переменной воздействия (количеством свободных мест) в авиационной индустрии обосновывается логикой спроса и предложения. Имеется экономическая

интуиция о том, что изменение предложения (количества свободных мест) может влиять на цену (спрос) на услугу (авиабилеты). При уменьшении количества свободных мест цены на билеты могут возрастать, так как спрос остается стабильным или растет, в то время как предложение уменьшается. Напротив, при наличии большого количества свободных мест цены могут снижаться для стимулирования спроса и максимизации заполненности рейсов. Подтверждение наших предположений можно найти в литературе. В статье Escobari (2012) "Dynamic Pricing, Advance Sales, and Aggregate Demand Learning in Airlines" [1] причинно-следственная связь между ценой авиабилета и количеством свободных мест обосновывается тем, что авиакомпании меняют цены на билеты в зависимости от текущего уровня бронирования и ожидаемого спроса. То есть с уменьшением количества свободных мест цены на оставшиеся билеты увеличиваются. Кроме того, авиакомпании стремятся максимизировать доходы, повышая цены по мере уменьшения доступности мест, учитывая поведение и ожидания потребителей.

В статье Bilotkach (2006) "Understanding Price Dispersion in the Airline Industry: Capacity Constraints and Consumer Heterogeneity" [2] автор подробно рассматривает механизмы, посредством которых количество свободных мест влияет на цену авиабилета. Основные аргументы, подтверждающие причинно-следственную связь и использующиеся в данной статье, можно резюмировать следующим образом: во-первых, авиакомпании используют динамическое ценообразование, чтобы максимизировать доходы. По мере приближения даты вылета и уменьшения количества свободных мест, компании повышают цены, чтобы извлечь максимальную выгоду от оставшихся мест. Это объясняется тем, что оставшиеся билеты становятся дефицитными, и компания может назначить за них более высокую цену; во-вторых, потребители отличаются по своей готовности платить и гибкости в планировании. Авиакомпании могут разделять рынок на различные сегменты (например, бизнес-путешественники и туристы) и устанавливать разные цены в зависимости от спроса в каждом сегменте. Когда количество свободных мест уменьшается, компании ориентируются на более платежеспособных потребителей, которые готовы заплатить больше за оставшиеся места; в-третьих, в статье обсуждается, что потребители с низкой эластичностью спроса (например, бизнес-путешественники) менее чувствительны к изменениям цены. Авиакомпании используют эту информацию для повышения цен на оставшиеся билеты по мере их уменьшения.

4. Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа.

Основные выводы из статьи Escobari, D. "Dynamic Pricing, Advance Sales, and Aggregate Demand Learning in Airlines" (2012) [1]:

1. Цены на авиабилеты изменяются в зависимости от времени до вылета и количества свободных мест. Меньше свободных мест – выше цены.
2. Авиакомпании используют данные о предварительных продажах для прогнозирования спроса и адаптации цен.
3. Цель – максимизация доходов через гибкое управление ценами.

В данной работе для прогнозирования цен и спроса на авиабилеты использовались линейные модели, однако такая модель может не учитывать нелинейные эффекты и взаимодействия между переменными. Также в работе использовалось динамическое моделирование, в целом, можно назвать данный инструмент хорошим, поскольку модель динамического ценообразования гибкая и позволяет учитывать временные изменения в спросе и предложении. Исследование проводилось на панельных данных, позволяет учитывать временные и индивидуальные эффекты, что увеличивает точность анализа.

Основные выводы из статьи Bilotkach (2006) "Understanding Price Dispersion in the Airline Industry: Capacity Constraints and Consumer Heterogeneity"[2]:

1. Ограничения по вместимости: Цены растут при уменьшении свободных мест.
2. Гетерогенность потребителей: Различия в готовности платить позволяют использовать ценовую дискриминацию.
3. Оптимизация доходов: Модели динамического ценообразования учитывают спрос и количество свободных мест.

В данном исследовании использовались эконометрические модели. Стоит отметить, что используемые модели достаточно гибки для учета различных факторов, влияющих на цену билетов. Тем не менее, модели могут быть ограничены в учете нелинейных эффектов и взаимодействий между переменными. В данной работе исследование также проводилось на панельных данных. Использование панельных данных позволяет учитывать временные и индивидуальные эффекты, что повышает точность результатов. Однако сбор и обработка таких данных требуют значительных ресурсов и могут содержать ошибки измерения.

5. 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.

Контрольные переменные:

Days_Before – количество дней перед датой рейса, обычно билеты на рейсы, купленные заблаговременно, имеют более низкие цены, чем те, которые покупаются ближе к дате вылета.

Distance – расстояние (км), обычно более дальние перелеты имеют более высокие цены из-за большего количества потребляемых ресурсов, таких как топливо и время, и, как следствие, стоимость обслуживания самолета становится более высокой.

Connection – прямой рейс (dummy, 1 - прямой рейс, то есть без пересадок, 0 - с пересадками), обычно прямые рейсы имеют более высокие цены из-за их привлекательности для пассажиров (в связи с меньшим временем полета) и удобства.

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

Инструментальная переменная - время вылета в “час пик” (dummy, 1 - время вылета в час пик, 0 - иначе), Hours.

Ненаблюдаемые переменные, порождающие эндогенность - Population – средняя численность населения в городах отправления (тыс./чел.).

Средняя численность населения может влиять на спрос на рейсы, что, в свою очередь, влияет на количество свободных мест. Высокий спрос в густонаселенных пунктах ведет к снижению числа свободных мест. Подобранная для борьбы с эндогенностью инструментальная переменная, отвечающая за вылет в “час пик”, должна быть валидной и релевантной. Данная переменная нам подходит, поскольку, во-первых, время вылета в “час пик” влияет на количество свободных мест (Last_ticket), так как рейсы в часы пик обычно пользуются большим спросом, то есть инструмент релевантен. Во-вторых, время вылета не должно напрямую влиять на ошибки и цену авиабилета (Total_Price), кроме как через количество свободных мест, особенно если мы контролируем такие переменные, как количество дней до вылета, расстояние и тип рейса (прямой или с пересадками), то есть инструмент валидный.

2 Генерация и предварительная обработка данных

1. Опишите математически предполагаемый вами процесс генерации данных.

Days_Before: Моделируется пуассоновским распределением P(60), так как количество дней до вылета имеет дискретное распределение с типичными значениями, которые можно предсказать.

Distance: Моделируется нормальным распределением с математическим ожиданием $\mu=1000$ и стандартным отклонением $\sigma=300$. Нормальное распределение подходит, так как расстояния между городами распределяются примерно нормально вокруг среднего значения.

Connection: Поскольку Connection является бинарной переменной, принимающей значения 0 и 1 , то она имеет распределение Бернулли $City \sim Ber(p)$. Предположим, что количество прямых рейсов составляет 40% от всех рейсов , откуда $p=0.4$.

Population: Моделируется на основе распределения Стьюдента с 5 степенями свободы, чтобы учесть потенциальные выбросы и тяжелые хвосты в данных о численности населения городов, что часто наблюдается в реальных демографических данных.

Hours: Для того, чтобы сгенерировать бинарную переменную как функцию от других переменных, необходимо сперва предположить форму условных вероятностей. Мы предположили, что условная вероятность времени вылета в "час пик" положительно связана с расстоянием и типом рейса (прямой или с пересадками) и отрицательно с количеством дней до вылета:

$$\mathrm{P}\left(\mathrm{Hours}_i = 1 | \mathrm{Days\_Before}_i, \mathrm{Distance}_i, \mathrm{Connection}_i\right)$$
$$= \Phi\left(\underbrace{\frac{0.4 \times \mathrm{Distance}_i - 0.3 \times \mathrm{Days\_Before}_i + 0.1 \times \mathrm{Connection}_i}{100} - 3.8}_{\text{индекс}}\right)$$, где Ф() - функция распределения стандартного нормального распределения.

Для удобства генерации мы также предполагаем, что условные вероятности инструментальной переменной зависят лишь от контрольных переменных.

Индекс был сгенерирован таким образом, чтобы обеспечить приемлемую дисперсию (1.44) и адекватную долю единиц в итоговой переменной (0.5074), что подтверждает адекватность модели.

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

$$P(\mathrm{Last\_ticket}_i = 1 | \mathrm{Days\_Before}_i, \mathrm{Distance}_i, \mathrm{Connection}_i, \mathrm{Population}_i, \mathrm{Hours}_i) =$$
$$= F_{\mathrm{Logistic}}\left(\ln(\mathrm{Days\_Before}_i + 1) + 0.05 \times \sqrt{\mathrm{Distance}_i} - 2 \times \mathrm{Connection}_i + 2 \times \ln(\mathrm{Population}_i + 1) + 3 \times \mathrm{Hours}_i - 18.5\right),$$

где F_{Logistic} - функция распределения стандартного логистического распределения.

Total_Price: Для генерации данной переменной мы выдвигаем следующую основную идею: влияние расстояния, количества дней до рейса, типа рейса и численности населения выше, когда количество свободных мест меньше 20.

Таким образом, получаем

1) Уравнение цены авиабилета при вылете вне часа пик:

$$\mathrm{Total\_Price}_{0i} = \underbrace{\underbrace{0.1 \times \ln(\mathrm{Population}_i + 1)}_{g_0^{\mathrm{unobs}}} + \underbrace{0.05 \times \sqrt{\mathrm{Distance}_i} + 3 \times \ln(\mathrm{Days\_Before}_i + 1) - 5 \times \mathrm{Connection}_i}_{g_0^{\mathrm{obs}}}}_{g_0}$$
$$+ \varepsilon_{0i}, \text{ где } \varepsilon_{0i} \sim (8 \times t(8))$$

2) Уравнение цены авиабилета при вылете в час пик:

$$\mathrm{Total\_Price}_{1i} = \underbrace{\underbrace{0.15 \times \ln(\mathrm{Population}_i + 1)}_{g_1^{\mathrm{unobs}}} + \underbrace{0.1 \times \sqrt{\mathrm{Distance}_i} + 4 \times \ln(\mathrm{Days\_Before}_i + 1) - 4 \times \mathrm{Connection}_i}_{g_1^{\mathrm{obs}}}}_{g_1}$$
$$+ \varepsilon_{1i}, \text{ где } \varepsilon_{1i} \sim (\mathrm{EXP}(10) - 10)$$

3) Наблюдаемая цена авиабилета:

$$\mathrm{Total\_Price}_i = \begin{cases} \mathrm{Total\_Price}_{1i}, \text{ если } \mathrm{Last\_Ticket}_i = 1 \\ \mathrm{Total\_Price}_{0i}, \text{ если } \mathrm{Last\_Ticket}_i = 0 \end{cases} =$$
$$= \mathrm{Total\_Price}_{1i} \times \mathrm{Last\_Ticket}_i + \mathrm{Total\_Price}_{0i} \times (1 - \mathrm{Last\_Ticket}_i)$$

2. Кратко обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

Рассмотрим индекс для переменной воздействия Last_Ticket и ее взаимосвязь с остальными переменными:

- ***Days_Before***. Положительная связь: чем больше дней остается до вылета, тем больше вероятность того, что еще остаются свободные места (то есть

переменная воздействия будет равна 0). Логарифм используется для смягчения эффекта больших значений.

- ***Distance***. Положительная связь: более дальние рейсы могут иметь меньше свободных мест из-за ограниченной вместимости самолета или повышенного спроса на более длинные перелеты.
- ***Connection***. Отрицательная связь: на прямых рейсах (то есть значение бинарной переменной равно 1) вероятность наличия свободных мест ниже по сравнению с рейсами с пересадками. Это связано с тем, что прямые рейсы более популярны и быстрее заполняются.

Рассмотрим уравнение для целевой переменной Ticket_Price и ее взаимосвязь с остальными переменными:

- ***Days_Before***. Логарифмическое преобразование количества дней до рейса сглаживает эффект уменьшения цены по мере увеличения количества дней до вылета, также коэффициент при переменной положительный, это указывает на то, что чем больше времени остается до вылета, тем выше цена билета.. При меньшем количестве свободных мест влияние этой переменной выше (=4), что может отражать повышенный спрос на билеты, купленные заранее.
- ***Distance***. Корень квадратный от расстояния используется для учета нелинейного эффекта расстояния на цену. Влияние этой переменной положительно, это значит, что с увеличением расстояния между городами цена билета также увеличивается. Это связано с тем, что большие расстояния требуют больше ресурсов (топливо и т.д.), что увеличивает стоимость рейса. Когда количество свободных мест меньше 20, влияние расстояния на цену выше (=0.1), таким образом отражается высокий спрос на дальние рейсы с ограниченным количеством мест.
- ***Connection***. Переменная Connection имеет значение 1 для прямых рейсов и 0 для рейсов с пересадками. Влияние этой переменной отрицательно, это означает, что прямые рейсы дешевле по сравнению с рейсами с пересадками. Это может быть связано с тем, что пересадки требуют дополнительных затрат на логистику и ресурсы, что увеличивает стоимость рейса.. При меньшем количестве свободных мест влияние этой переменной немного меньше (=-4), что может отражать высокий спрос на все типы рейсов.

3. 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 ни для одной из бинарных переменных.

Таким образом, было сгенерировано 100 000 наблюдений и построена корреляционная матрица (см. Таблица 1). Кроме того, были сформированы таблицы с описательными

статистиками для непрерывных(см. Таблица 2) и бинарных переменных(см. Таблица 3), таблица для бинарных переменных подтверждает то, что необходимые указания соблюдены: доли й в бинарных переменных лежат в диапазоне от 0.4 до 0.5.

Таблица 1. Корреляционная матрица.

	Ticket_Price	**Last_Ticket**	**Distance**	**Days_Before**	**Connection**	**Hours**
Ticket_Price	1	0.906	0.335	0.078	-0.598	0.313
Last_Ticket	0.906	1	0.279	0.037	-0.637	0.335
Distance	0.335	0.279	1	0.001	0.007	0.613
Days_Before	0.078	0.037	0.001	1	-0.005162	-0.016
Connection	-0.598	-0.637	0.007	-0.005	1	0.003
Hours	0.313	0.335	0.613	-0.016	0.003	1

Таблица 2. Описательная статистика для непрерывных переменных.

	Ticket_Price	**Distance**	**Days_Before**
mean	255.362	1002.081	59.992
std	33.020	299.585	7.736
50%	263.015	1001	60
min	110.434	24	27
max	381.908	2428	96

Таблица 3. Описательная статистика для бинарных переменных.

	Количество "1"	**Доля "1"**

Last_Ticket	51743	0.517
Connection	40129	0.401
Hours	50740	0.507

4. Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20% до 30% наблюдений.
В результате деления выборки тестовая выборка включает 20% наблюдений от общего количества данных. Таким образом, в тестовую выборку вошло 20 000 наблюдений, а в обучающую - 80 000 наблюдений.

3 Классификация

0. В каждом из заданий, если не сказано иного, необходимо использовать хотя бы 3 (на ваш выбор) из следующих методов: наивный Байесовский классификатор, метод ближайших соседей, случайный лес, градиентный бустинг и логистическая регрессия.
В данном пункте в нашей работе использовались следующие методы: метод ближайших соседей, случайный лес, градиентный бустинг.
1. Отберите признаки, которые могут быть полезны при прогнозировании переменной воздействия и кратко обоснуйте выбор каждой из них. Не включайте в число этих признаков целевую переменную.
Для прогнозирования переменной воздействия (Last_Ticket) были отобраны все 3 контрольных переменных и инструментальная переменная из следующих соображений об их взаимосвязи:

- Количество дней перед датой рейса (Days_Before) является важным фактором для прогнозирования количества свободных мест, поскольку чем ближе дата вылета, тем больше вероятность того, что билеты уже распроданы, то есть количество мест мало, особенно если это популярное направление или праздничный период.
- Расстояние (Distance) также может влиять на количество свободных мест. Для более дальних направлений обычно доступно меньше свободных мест из-за ограниченной вместимости самолета.
- Переменная Connection, отвечающая за факт того, является ли рейс прямым, также была отобрана, поскольку тип рейса и количество мест могут быть связаны между собой уровнем спроса на билеты. Если есть возможность совершить прямой рейс, пассажиры, скорее всего, предпочтут его, что может привести к меньшему количеству свободных мест на таких рейсах.
- Переменная Hours - время вылета в часы пика также может оказывать влияние на количество свободных мест на рейсе. В часы пика спрос на билеты может быть выше, что может привести к уменьшению количества свободных мест.

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).
Предварительно данные были нормализованы при помощи StandardScaler. Рассмотрим каждую модель и ее результаты по порядку:

1. Метод ближайших соседей с гиперпараметром n_neighbors = 5 показал хорошую точность: точность на обучающей выборке равна 0.865; на тестовой выборке точность равна 0.8; кросс-валидационная точность на обучающей выборке составила 0.799.
2. Случайный лес (со следующими гиперпараметрами: max_depth = 16, max_features = 'sqrt', random_state = 22) также показал хорошую точность: точность на обучающей выборке равна 0.879; на тестовой выборке точность равна 0.809; кросс-валидационная точность на обучающей выборке составила 0.81.
3. Градиентный бустинг с параметром random_state = 22 также показал хорошую точность: точность на обучающей выборке равна 0.823; на тестовой выборке точность равна 0.820; кросс-валидационная точность на обучающей выборке составила 0.821.

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.
Для выбранных 3 методов были подобраны гиперпараметры при помощи GridSearchCV. Для всех методов точность ACC на тестовой выборке и кросс-валидационная точность на обучающей выборке возросли. Рассмотрим все методы по порядку:

1. *Метод ближайших соседей.* В результате были подобраны следующие гиперпараметры: {'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}. Подбор гиперпараметров с использованием кросс-валидации позволил улучшить качество модели ближайших соседей. Небольшое улучшение точности и на кросс-валидации, и на тестовой выборке свидетельствует о том, что подобранные гиперпараметры способствуют стабильности модели без явных признаков переобучения. (см. Таблицу 4) Изменение степени расстояния Минковского с p=2 (евклидовое расстояние) на p=1 (Манхэттенское расстояние)

указывает на то, что модель лучше классифицировала данные, когда использовалась сумма абсолютных различий вместо квадратов различий.

Таблица 4. Точность параметров для метода ближайших соседей до и после кросс-валидации.

Параметры	Изначальные значения	Подобранные значения
Гиперпараметры	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 5, 'p': 2, 'weights': 'uniform'}	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}
Кросс-валидационная точность на обучающей выборке	0.799	0.806
Точность на тестовой выборке	0.8	0.806

2. *Случайный лес.* В результате были подобраны следующие гиперпараметры: {'max_depth': 7, 'max_features': 'sqrt', 'n_estimators': 200, 'random_state': 22}. Подбор гиперпараметров с использованием кросс-валидации позволил улучшить качество модели случайного леса. Небольшое улучшение точности и на кросс-валидации, и на тестовой выборке свидетельствует о том, что подобранные гиперпараметры способствуют стабильности модели без явных признаков переобучения. (см. Таблицу 5) Уменьшение максимальной глубины деревьев с 16 до 7 указывает на то, что более простые деревья с меньшей глубиной помогают избежать переобучения и лучше обобщают данные, а увеличение количества деревьев в лесу с 100 до 200 приводит к более точной оценке модели и уменьшению случайных колебаний, что также способствует улучшению производительности модели.

Таблица 5. Точность параметров для случайного леса до и после кросс-валидации.

Параметры	Изначальные значения	Подобранные значения
Гиперпараметры	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': 22, 'verbose': 0, 'warm_start': False}	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 7, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 22, 'verbose': 0, 'warm_start': False}

Кросс-валидационная точность на обучающей выборке	0.810	0.822
Точность на тестовой выборке	0.809	0.822

3. *Градиентный бустинг.* В результате были подобраны следующие гиперпараметры: {'max_depth': 5, 'max_features': None, 'n_estimators': 100, 'random_state': 22}. Кросс-валидационная точность на обучающей выборке улучшилась с 0.821 до 0.822. Это показывает, что подобранные гиперпараметры обеспечили незначительное улучшение модели на обучающих данных. В то же время точность на тестовой выборке немного снизилась с 0.821 до 0.820. Это указывает на то, что небольшое увеличение сложности модели (Увеличение максимальной глубины деревьев с 3 до 5 уровней, то есть создание более глубоких деревьев) не привело к значительным изменениям в ее общей производительности на тестовых данных, что свидетельствует о стабильности модели. (см. Таблицу 6)

Таблица 6. Точность параметров для градиентного бустинга до и после кросс-валидации.

Параметры	**Изначальные значения**	**Подобранные значения**
Гиперпараметры	{'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'log_loss', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': 22, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}	{'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'log_loss', 'max_depth': 5, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': 22, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}
Кросс-валидационная точность на обучающей выборке	0.821	0.822
Точность на тестовой выборке	0.821	0.820

<u>Повышенная сложность:</u> 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 преимущество OOB ошибки по сравнению с кросс-валидацей.

Кросс-валидация, хотя и является более универсальной техникой, требует больше времени на выполнение, так как модель обучается и оценивается несколько раз на различных подвыборках данных. При использовании OOB ошибки гиперпараметры модели могут быть настроены быстрее. OOB ошибка автоматически предоставляет оценку на основе наблюдений, не участвовавших в обучении конкретного дерева, исключая необходимость многократного деления данных на фолды и их пересчета, как это требуется при кросс-валидации. Кроме того, преимуществом OOB ошибок является то, что все данные используются и для обучения, и для оценки, тогда как в кросс-валидации часть данных выделяется в качестве валидационной выборки, что уменьшает количество данных, доступных для обучения в каждом цикле.
Тем не менее, на наших данных модель случайного леса с использованием OOB ошибок показала немного меньшую точность, чем с использованием кросс-валидации. (см. Таблицу 7)

Таблица 7. Точность параметров для случайного леса с использованием кросс-валидации и OOB ошибок.

Параметры	**Кросс-валидация**	**OOB ошибки**
Гиперпараметры	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 7, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 22, 'verbose': 0, 'warm_start': False}	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': None, 'oob_score': True, 'random_state': 22, 'verbose': 0, 'warm_start': False}
Кросс-валидационная точность на обучающей выборке	0.822	0.817
Точность на тестовой выборке	0.822	0.817

4. Повторите предыдущий пункт, используя любой альтернативный критерий качества модели. Обоснуйте возможные преимущества и недостатки этого альтернативного критерия.

В качестве альтернативного критерия была выбрана F-1 мера - гармоническое среднее между точностью и полнотой. Использование F1-меры в качестве критерия качества в GridSearchCV имеет свои преимущества, особенно в ситуациях с

несбалансированными классами. Она позволяет получить баланс между точностью и полнотой. Однако недостатком является то, что F-1 мера может быть менее интерпретируемой по сравнению с отдельными метриками точности и полноты.
В результате использования данного критерия для кросс-валидации на обучающей выборке были получены абсолютно идентичные оптимальные гиперпараметры. Соответственно, результаты точности модели останутся теми же, что и в задании 3.3. То, что оптимальные значения гиперпараметров не поменялись, можно объяснить тем, что классы в данных сбалансированы. Также алгоритм подбора гиперпараметров в GridSearchCV может приводить к тому, что даже при изменении метрики оценки, он все равно выбирает те же оптимальные значения гиперпараметров, если они обеспечивают лучшее общее качество модели.

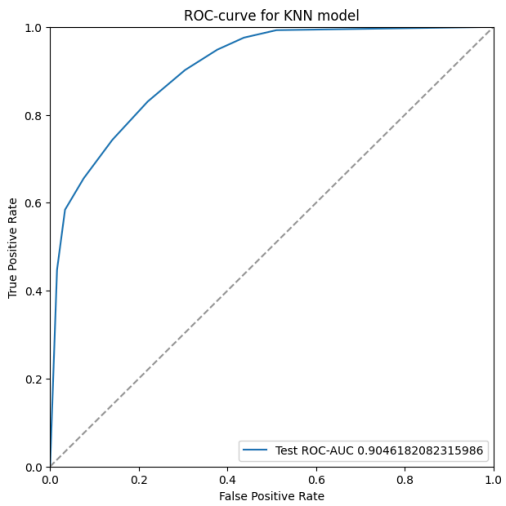
<u>Повышенная сложность:</u> дополнительно самостоятельно запрограммируйте не представленный в стандартных библиотеках критерий качества и используйте его для тюнинга гиперпараметров. Сравните результат стандартного и вашего критериев.

Для выполнения данного задания была создана функция, которая вычисляет среднее арифметическое между точностью и полнотой модели. Это простой, но интуитивно понятный критерий качества, который учитывает как долю правильно классифицированных положительных результатов (точность), так и долю всех действительно положительных наблюдений, которые были правильно определены (полнота).
Однако использование данного критерия с нашими данными привело к небольшому уменьшению точности всех моделей. Если изменение критерия привело к уменьшению точности, это может означать, что модель стала более склонной к снижению ложноотрицательных прогнозов за счет увеличения ложноположительных. Это может произойти, например, если модель стала более чувствительной к выбросам или шуму в данных.

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

Для метода ближайших соседей AUC получился равен 0.9046 (см. Рисунок 1), для случайного леса - 0.9184 (см. Рисунок 2), а для градиентного бустинга, чуть меньше, - 0.91796 (см. Рисунок 3). AUC очень близок к 1 для всех методов, это говорит об отличном качестве моделей, то есть модели одинаково хорошо “научились” работать как с положительными, так и с отрицательными примерами при существующем в обучающей выборке балансе классов.

Рисунок 1. ROC-кривая для метода ближайших соседей. **Рисунок 2. ROC-кривая для случайного леса.**



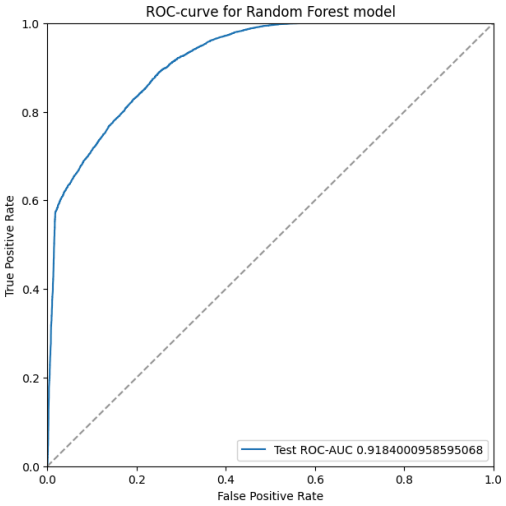
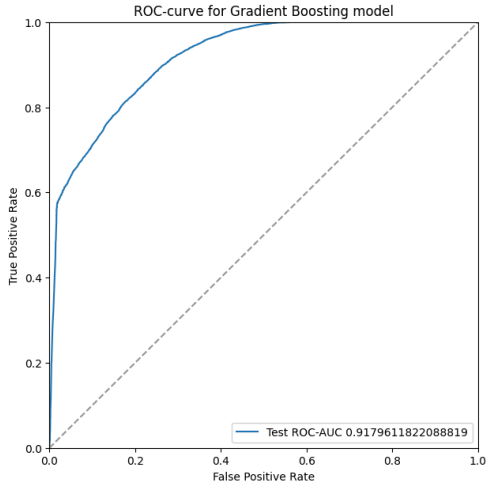


Рисунок 3. ROC-кривая для градиентного бустинга.



6. Постройте матрицу путаницы 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.

Для различных видов прогнозов цены были заданы следующим образом: при правильной классификации положительного случая (TP) модель приносит 100 единиц прибыли; при ошибочной классификации отрицательного случая как положительного (FP), возникает убыток в 50 единиц; при правильной классификации отрицательного случая (TN) прибыль или убыток отсутствуют; при ошибочной классификации положительного случая как отрицательного (FP) , возникает наибольший убыток в 200 единиц. Такой выбор цен обусловлен необходимостью максимально учитывать ошибки классификации, которые могут быть дорогостоящими для бизнеса. Высокий штраф за ложноположительные и ложноотрицательные прогнозы (особенно за ложноотрицательные, которых по построенным матрицам путаницы оказалось больше) стимулирует модели снижать количество таких ошибок, что важно для максимизации прибыли.

В таблице (см. Таблицу 8) представлены результаты работы трех различных методов машинного обучения: метод ближайших соседей, случайный лес и градиентный бустинг. Для каждого метода были рассчитаны AUC, оптимальный порог прогнозирования и прибыль на тестовой выборке.

Таблица 8. AUC, порог и прибыль для методов классификации.

Model	AUC (Test)	Optimal Threshold	Profit (Test)
K-Nearest Neighbors	0.9046	0	1041300
Random Forest	0.9184	0	1041300
Gradient Boosting	0.9180	0.000632	1041300

Все три модели показывают одинаковую прибыль на тестовой выборке, равную 1041300. Это может свидетельствовать о том, что при данных условиях задачи и ценах различных видов прогнозов, все модели показывают схожую эффективность в максимизации прибыли. Стоит отметить, что для метода ближайших соседей и случайного леса оптимальный порог прогнозирования оказался равен 0. Это может указывать на то, что модели хорошо справляются с задачей классификации даже при нулевом пороге. Для градиентного бустинга оптимальный порог прогнозирования оказался равен 0.000632. Это очень низкое значение, что может говорить о высокой чувствительности модели к малым изменениям вероятностей.

<u>Повышенная сложность:</u> предложите, содержательно обоснуйте и примените собственную, отличную от линейной функцию прибыли от прогнозов.

Была предложена следующая функция прибыли:

$$100 \times \sqrt{TP} - (50 + \text{marketing cost}) \times FP^{1.5} - (200 + \text{retention cost}) \times \ln(FN + 1),$$

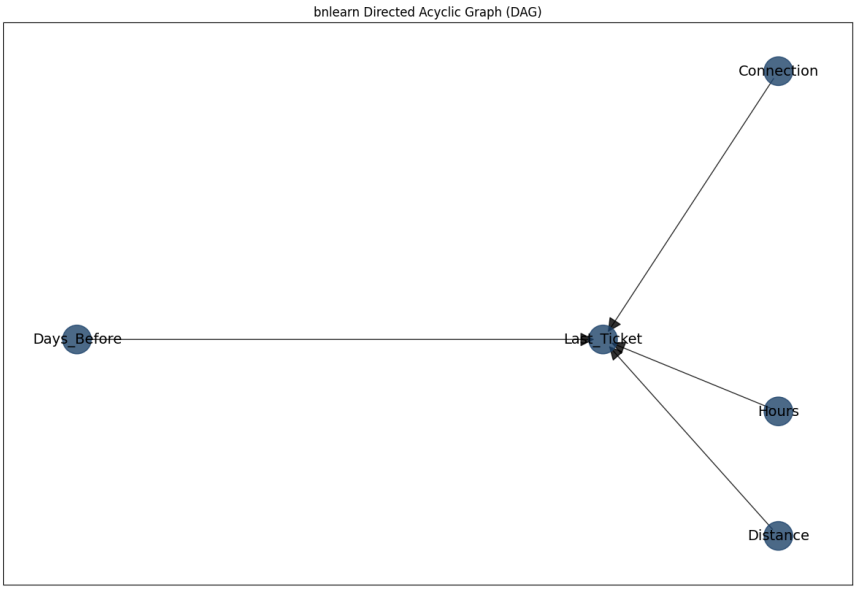
где marketing cost - маркетинговые расходы, retention cost - расходы на удержание. Использование квадратного корня для TP отражает убывающую отдачу от правильных предсказаний положительного события. Например, в маркетинге привлечение первых нескольких клиентов может быть более ценно, чем привлечение каждого последующего клиента: начальное увеличение точных положительных прогнозов значительно улучшает результаты, но дальнейшее улучшение приносит все меньше и меньше дополнительных выгод. Использование степенной функции для FP означает, что убытки от ложных положительных предсказаний увеличиваются быстрее, чем линейно, с ростом их количества, то есть если ложные предсказания приводят к ненужным маркетинговым затратам, то при большом количестве таких ошибок общий ущерб увеличивается больше, чем пропорционально. Кроме того, убытки от ложных положительных предсказаний включают маркетинговые расходы, поскольку, когда модель ошибочно предсказывает, что клиент совершит покупку (FP), компания может потратить ресурсы на маркетинговые кампании, направленные на этого клиента. Эти кампании могут включать рекламу, персонализированные предложения и другие маркетинговые активности. Если эти клиенты в конечном итоге не покупают продукт,

компания теряет деньги, потраченные на эти маркетинговые ходы. Использование логарифма для FN отражает быстрое увеличение убытков при малых значениях FN и замедление роста убытков при больших значениях FN. Таким образом, если упущенные возможности приводят к потере потенциальных клиентов, то начальные упущения могут быть более болезненными и критичными, чем последующие. Кроме того, убытки от ложных отрицательных предсказаний включают расходы на удержание, поскольку, когда модель ошибочно предсказывает, что клиент не заинтересован (FN), компания упускает возможность удержать этого клиента или сделать предложение, которое могло бы привести к продаже. Удержание клиентов может включать программы лояльности, специальные скидки или другие мероприятия, направленные на повышение удовлетворенности и повторные покупки. Неспособность идентифицировать потенциальных клиентов приводит к упущенной прибыли и дополнительным расходам на удержание.

7. Опишите предполагаемые связи между переменными в форме ориентированного ациклического графа (DAG). Обучите структуру Байесовской сети на обучающей выборке и сравните точность прогнозов вашего и обученного DAG на тестовой выборке.

С помощью пакета bnlearn был сформирован DAG. (см. Рисунок 4)

Рисунок 4. Ориентированный ациклический граф.



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

Лучшей моделью является случайный лес с подобранными оптимальными гиперпараметрами, этот метод имеет самое высокое значение AUC, ACC на тестовой выборке и кросс-валидационную точность на обучающей выборке, что означает отличное качество модели и способность различать между положительными и отрицательными классами. Поскольку прибыль на тестовой выборке одинаковая для всех моделей, выбор модели с наивысшим AUC является наиболее логичным.

Худшей моделью является метод ближайших соседей с изначальными гиперпараметрами, так как этот метод имеет самое низкое значение ACC на тестовой выборке и кросс-валидационную точность на обучающей выборке.

9. Повышенная сложность: включите в анализ дополнительный метод классификации, не рассматривавшийся в курсе и не представленный в библиотеке scikit-learn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов.

В качестве дополнительного метода был выбран CatBoostClassifier. CatBoost - это алгоритм градиентного бустинга, разработанный Yandex, который использует симметричную оптимизацию с использованием градиентов второго порядка. Он построен на основе принципов классического алгоритма градиентного бустинга, но с добавлением дополнительных функций, таких как обработка категориальных признаков, поддержка отсутствующих значений и возможность работать с большими наборами данных. Кроме этого, CatBoost использует симметричную оптимизацию, которая учитывает как положительные, так и отрицательные градиенты, что позволяет строить более точные деревья. Данный алгоритм обеспечивает высокую производительность и эффективное предотвращение переобучения.

Преимущества CatBoost:

1. Обработка категориальных признаков: CatBoost может обрабатывать как числовые, так и категориальные признаки без необходимости предварительного преобразования. Он использует встроенный алгоритм, известный как Count Encoding Trick (CET), который преобразует категории в бинарные признаки.
2. Поддержка отсутствующих значений: CatBoost умеет обрабатывать отсутствующие значения. Он присваивает отсутствующим значениям преобладающий целочисленный признак или среднее значение вещественного признака.
3. Эффективность и масштабируемость: CatBoost оптимизирован для работы с большими наборами данных. Его распределенный алгоритм позволяет распараллеливать процесс обучения на нескольких машинах.

Недостатки:

1. Требовательность к ресурсам: CatBoost может потреблять значительное количество оперативной памяти и вычислительных ресурсов, особенно при работе с большими наборами данных.
2. Чувствительность к гиперпараметрам: Как и другие алгоритмы бустинга, CatBoost требует тщательной настройки гиперпараметров для достижения оптимальной производительности.

4 Регрессия

0. В каждом из заданий, если не сказано иного, необходимо использовать хотя бы 3 (на ваш выбор) из следующих методов: случайный лес, метод наименьших квадратов, метод ближайших соседей и градиентный бустинг.

В данном пункте в нашей работе использовались следующие методы: метод ближайших соседей, случайный лес, градиентный бустинг.

1. Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия.

Для прогнозирования целевой переменной (Ticket_Price) были отобраны все 3 контрольных переменных, исходя из следующих соображений об их взаимосвязи:

- Количество дней до вылета (Days_Before) является важным фактором, влияющим на цену авиабилета. Обычно, чем ближе дата вылета, тем выше цена билета. Авиакомпании часто повышают цены на билеты по мере приближения даты рейса, чтобы максимизировать доходы.
- Расстояние между пунктами отправления и назначения (Distance) напрямую влияет на стоимость авиаперелета. Более длинные рейсы требуют большего расхода топлива и других ресурсов, что приводит к увеличению цены билета.
- Наличие пересадок (Connection) может существенно влиять на цену авиабилета. Прямые рейсы, как правило, стоят дороже, чем рейсы с пересадками, так как они удобнее и требуют меньшего времени в пути.

2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов с помощью RMSE и MAPE:

• на обучающей выборке.

• на тестовой выборке.

• с помощью кросс-валидации (используйте только обучающую выборку).

Предварительно данные были нормализованы при помощи StandardScaler. Стоит отметить, что при использовании cross_val_score для расчета кросс-валидационной точности был использован критерий neg_root_mean_squared_error, который возвращал значения метрик с отрицательным знаком. Это сделано для того, чтобы метрики соответствовали стандарту "чем больше, тем лучше" (высокие значения показывают лучшую модель). Поскольку RMSE (корень из среднеквадратичной ошибки) является метрикой ошибки и, по своей природе, чем меньше значение, тем лучше, отрицательное значение RMSE инвертирует эту логику. Таким образом, полученные значения для верной интерпретации были взяты по модулю. Рассмотрим каждую модель и ее результаты по порядку:

1. Метод ближайших соседей с гиперпараметром n_neighbors = 5 показал хорошую точность: RMSE и MAPE на обучающей выборке равны 20.971 и 0.064%; на тестовой выборке 44.457 и 0.152%; кросс-валидационные RMSE и MAPE на обучающей выборке составили 25.703 и 0.078%.
2. Случайный лес (со следующими гиперпараметрами: n_estimators = 100, random_state = 22) также показал хорошую точность: RMSE и MAPE на обучающей выборке равны 16.429 и 0.047%; на тестовой выборке 42.103 и 0.144%; кросс-валидационные RMSE и MAPE на обучающей выборке составили 27.081 и 0.082%.
3. Градиентный бустинг (со следующими гиперпараметрами: n_estimators = 100, random_state = 22) также показал хорошую точность: RMSE и MAPE на обучающей выборке равны 23.385 и 0.071%; на тестовой выборке 44.148 и

0.157%; кросс-валидационные RMSE и MAPE на обучающей выборке составили 23.486 и 0.071%.

3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите 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.

Для выбранных 3 методов были подобраны гиперпараметры при помощи GridSearchCV. Для двух методов RMSE на тестовой выборке и кросс-валидационная на обучающей выборке снизились. Рассмотрим все методы по порядку:

1. *Метод ближайших соседей.* В результате были подобраны следующие гиперпараметры: {'n_neighbors': 10, 'p': 1, 'weights': 'uniform'}. Подбор гиперпараметров с использованием кросс-валидации позволил улучшить качество модели ближайших соседей. Подобранные значения гиперпараметров дали небольшое улучшение на кросс-валидационной выборке (25.703 против 24.616), что указывает на лучшее общее поведение модели на обучающей выборке. Значительное улучшение RMSE на тестовой выборке (с 44.457 до 32.382) при использовании подобранных гиперпараметров указывает на то, что модель стала значительно лучше обобщать на новых данных. Смена метрики с евклидовой (p=2) на манхэттенскую (p=1) могла улучшить производительность модели. Также увеличение числа соседей с 5 до 10 помогает модели быть менее чувствительной к шуму и выбросам. (см. Таблицу 9)

Таблица 9. Точность параметров для метода ближайших соседей до и после кросс-валидации.

Параметры	Изначальные значения	Подобранные значения
Гиперпараметры	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 5, 'p': 2, 'weights': 'uniform'}	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 10, 'p': 1, 'weights': 'uniform'}
Кросс-валидационная RMSE на обучающей выборке	25.703	24.616
RMSE на тестовой выборке	44.457	32.382

2. *Случайный лес.* В результате были подобраны следующие гиперпараметры: {'max_depth': 7, 'max_features': None, 'n_estimators': 200, 'random_state': 22}. Снижение RMSE на обучающей выборке свидетельствует о том, что

подобранные гиперпараметры лучше соответствуют структуре данных, уменьшая ошибку модели на обучающей выборке. Снижение RMSE на тестовой выборке указывает на улучшение обобщающей способности модели. Это означает, что модель с подобранными гиперпараметрами лучше предсказывает данные, которые не были использованы для обучения. Увеличение числа деревьев с 100 до 200 привело к улучшению стабильности предсказаний и снижению дисперсии модели. Также Установление максимальной глубины дерева на уровне 7 позволило избежать переобучения модели, сохраняя баланс между сложностью модели и ее способностью к обобщению. (см. Таблицу 10)

Таблица 10. Точность параметров для случайного леса до и после кросс-валидации.

Параметры	**Изначальные значения**	**Подобранные значения**
Гиперпараметры	{'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': None, 'max_features': 1.0, 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': 22, 'verbose': 0, 'warm_start': False}	{'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': 7, 'max_features': None, 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 22, 'verbose': 0, 'warm_start': False}
Кросс-валидационная RMSE на обучающей выборке	27.081	23.495
RMSE на тестовой выборке	42.103	38.596

3. *Градиентный бустинг.* В результате оптимальные гиперпараметры оказались такими же, как и изначальные: {'max_depth': 3, 'max_features': None, 'n_estimators': 100, 'random_state': 22}. (см. Таблицу 11) Результаты показывают, что начальные гиперпараметры модели градиентного бустинга были уже оптимальными для данной задачи, что объясняет отсутствие улучшений после подбора гиперпараметров.

Таблица 11. Точность параметров для градиентного бустинга до и после кросс-валидации.

Параметры	**Изначальные значения**	**Подобранные значения**
Гиперпараметры	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'squared_error', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': 22,	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'squared_error', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': 22,

	'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}	'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}
Кросс-валидационная точность на обучающей выборке	23.486	23.486
Точность на тестовой выборке	44.148	44.148

<u>Повышенная сложность:</u> подберите на обучающей выборке оптимальные значения гиперпараметров градиентного бустинга ориентируясь на значение OOB (out-of-bag) ошибки. Сопоставьте гиперпараметры и точность на тестовой выборке для градиентного бустинга в зависимости от того, используется кросс-валидация или OOB ошибка.

Значения кросс-валидационной точности на обучающей выборке при использовании обоих методов подбора гиперпараметров практически одинаковы. (см. Таблицу 12) Это говорит о том, что модель хорошо подогнана к обучающим данным в обоих случаях. Значительное снижение RMSE на тестовой выборке при использовании OOB ошибки по сравнению с кросс-валидацией говорит о том, что модель, подобранная с использованием OOB ошибки, лучше обобщается на новые данные. Это означает, что модель, обученная с учетом OOB ошибки, обладает лучшей способностью к предсказанию на тестовой выборке.

Таблица 12. Точность параметров для градиентного бустинга для кросс-валидации и OOB ошибки.

Параметры	Кросс-валидация	OOB ошибки
Гиперпараметры	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'squared_error', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': 22, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'squared_error', 'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_iter_no_change': None, 'random_state': 22, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}
Кросс-валидационная точность на обучающей выборке	23.486	23.56
Точность на тестовой выборке	44.148	32.374

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

Среди всех рассмотренных моделей лучшей является градиентный бустинг с использованием OOB ошибки.

Причины:

- Подобранные гиперпараметры с использованием OOB ошибки показали наименьшее значение RMSE на тестовой выборке (32.3737).
- Кросс-валидационная RMSE на обучающей выборке также является одной из наименьших, что свидетельствует о хорошей способности модели к обобщению и предсказанию целевой переменной на новых данных.

Худшей моделью является градиентный бустинг без использования OOB ошибки с изначальными гиперпараметрами.

Причины:

- Изначальные гиперпараметры показали наибольшее значение RMSE на тестовой выборке (44.148), что свидетельствует о наихудшем качестве предсказаний.
- Разница между кросс-валидационной RMSE на обучающей выборке и RMSE на тестовой выборке указывает на возможное переобучение или недостаточную способность модели к обобщению. При том данная разница больше, чем у модели KNN с изначальными параметрами, именно поэтому, несмотря на то, что у KNN RMSE чуть больше, мы делаем вывод о том, что модель градиентного бустинга все же хуже.

5. Повышенная сложность: включите в анализ дополнительный метод регрессии, не рассматривавшийся в курсе и не представленный в библиотеке scikitlearn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов.

В качестве дополнительного метода был выбран CatBoostRegressor. Подробно об этом методе уже было написано выше в задании 3.9. RMSE на тестовой выборке для модели CatBoost получилось равно 33.983, а для лучшей модели (градиентный бустинг с использлванием OOB ошибки) - 32.373. Таким образом, модель градиентного бустинга все еще лучше.

5 Эффекты воздействия

1. Математически запишите и содержательно проинтерпретируйте потенциальные исходы целевой переменной. Объясните, как они связаны с наблюдаемыми значениями целевой переменной.

Потенциальные исходы целевой переменной Ticket_Price (цена авиабилета) для каждой единицы наблюдения зависят от воздействия переменной Last_Ticket (количество свободных мест). В нашем случае Last_Ticket = 1, если свободных мест меньше 20, и Last_Ticket = 0, если свободных мест больше 20.

Для каждого наблюдения (i) наблюдаемое значение целевой переменной определяется следующим образом:

$Ticket\_Price_i = Last\_Ticket_i * Ticket\_Price_{1i} + (1 - Last\_Ticket)_i * Ticket\_Price_{0i}$
Математически запишем и проинтерпретируем потенциальные исходы:
1. Средний Эффект Воздействия (ATE): $\mathrm{ATE} = \mathrm{E}\left(\mathrm{Ticket\_Price}_{1i} - \mathrm{Ticket\_Price}_{0i}\right)$
ATE показывает среднюю разницу в цене авиабилета, если для всех наблюдений изменить количество свободных мест с Last_Ticket = 0 на Last_Ticket = 1, то есть насколько в среднем изменится цена авиабилета, если изменить количество свободных мест.
2. Условный Средний Эффект Воздействия (CATE):
$\mathrm{CATE}_i = \mathrm{E}\left(\mathrm{Ticket\_Price}_{1i}|X_i\right) - \mathrm{E}\left(\mathrm{Ticket\_Price}_{0i}|X_i\right) = g_1(X_i) - g_0(X_i)$
,где X_i — набор контрольных переменных, таких как количество дней перед датой рейса, расстояние, тип рейса и т.д. CATE показывает среднюю разницу в цене авиабилета при изменении количества свободных мест, учитывая конкретные значения контрольных переменных.
3. Локальный Средний Эффект Воздействия (LATE):
$\mathrm{LATE} = \mathrm{E}(\mathrm{Ticket\_Price}_{1i} - \mathrm{Ticket\_Price}_{0i}|\mathrm{Last\_Ticket}_{1i} > \mathrm{Last\_Ticket}_{0i})$
Локальный средний эффект воздействия полезен в ситуациях, когда воздействие переменной является эндогенным, и мы используем инструментальную переменную (Hours) для оценки эффекта воздействия. LATE показывает средний эффект для тех рейсов, у которых изменение времени вылета действительно влияет на количество свободных мест.
Как же потенциальные исходы связаны с наблюдаемыми значениями целевой переменной? Наблюдаемые значения целевой переменной — это реализация одного из потенциальных исходов в зависимости от фактического значения переменной воздействия. Таким образом, наблюдаемая цена авиабилета Ticket_Price будет либо $Ticket\_Price_{1i}$, либо $Ticket\_Price_{0i}$, в зависимости от того, есть ли воздействие (меньше 20 свободных мест) или нет.

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. Результаты представьте в форме таблицы.

Используя формулы для ATE, LATE и CATE выше, были получены следующие результаты (см. Таблицу 13):

Таблица 13. ATE, LATE и CATE

Эффект	Значение
ATE	43.514
LATE	43.460
CATE	43.490

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.

С помощью данного подхода были получены следующие результаты (см. Таблицу 14):

Таблица 14. ATE и ATE как разница в средних

	ATE
Точная оценка с помощью потенциальных исходов	43.514
Наивная оценка с помощью наблюдаемых исходов	59.835

Как видно из таблицы, наивная оценка сильно больше. Дело в том, что оценка среднего эффекта воздействия (ATE) как разницы в средних по выборкам тех, кто получил и не получил воздействие, может иметь несколько недостатков в контексте формирования цены авиабилета в зависимости от количества свободных мест на борту:

- Гетерогенность эффекта: воздействие может различаться в зависимости от характеристик наблюдений (например, направления рейса, времени суток и т.д.). Средний эффект не отражает возможные различия в воздействии на разные группы пассажиров или рейсов. Например, влияние свободных мест на цену может быть разным для международных и внутренних рейсов, но разница в средних не позволит это учесть.
- Селективное смещение: выборка наблюдений не является случайной, а определена на основании некоторых критериев. Например, люди могут покупать билеты в разные моменты времени по разным причинам (например, бизнес-поездки, отпуск и т.д.), что влияет как на выбор (наличие свободных мест), так и на цену билета. Эти различия не учитываются при простом сравнении средних.
- Временные различия: цены на авиабилеты могут изменяться в зависимости от времени до вылета, общего уровня спроса и других временных факторов. Разница в средних не учитывает временные изменения, которые могут значительно исказить оценку эффекта воздействия.

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.

• 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.

Были получены результаты оценки ATE всеми перечисленными выше подходами и представлены в виде таблицы (см. Таблицу 15):

Таблица 15. Оценки ATE с использованием различных подходов

	Оценка

ATE	43.514
ATE naive	59.835
ATE ols	54.627
ATE S-learner	58.001
ATE_IPW	44.345
ATE_DR	59.974
ATE_dml_standard	61.5

Метод взвешивания на обратные вероятности (IPW) оказался наиболее близким к первоначальному ATE. В данном случае, это могло означать, что все важные переменные, влияющие на вероятность воздействия, были включены в модель, и она была хорошо обучена.

Сравнение результатов и ключевые предпосылки методов:

1. ATE naive (59.835): простой метод, который оценивает ATE как разницу в средних значениях между группами с воздействием и без воздействия.

Предпосылка: Отсутствие систематических различий между группами, помимо воздействия (отсутствие конфаундинга).

Проблемы: В реальных данных предпосылка часто нарушается из-за наличия конфликтующих переменных, что может приводить к смещенной оценке.

2. ATE OLS (54.627): метод наименьших квадратов (OLS), где воздействие включено в качестве объясняющей переменной вместе с другими контролируемыми переменными.

Предпосылка: Все релевантные переменные включены в модель, и между воздействием и ошибкой модели нет корреляции.

Проблемы: Если не все конфаундеры учтены, оценка может быть смещенной. В нашем случае, такие переменные как сезонные колебания или особенности рейсов могут не быть полностью учтенными.

3. ATE S-learner (58.001): метод S-learner строит модель предсказания исхода, включая воздействие в качестве одной из объясняющих переменных.

Предпосылка: Модель правильно специфицирована и включает все релевантные переменные.

Проблемы: Если модель неправильно специфицирована или не включает все важные переменные, оценка может быть смещенной.

4. ATE IPW (44.345): взвешивание на обратные вероятности (IPW), где каждое наблюдение взвешивается обратной вероятностью получения воздействия.

Предпосылка: Правильная спецификация модели вероятности воздействия (пропенсити скор).

Проблемы: Если модель propensity score неверно специфицирована, оценки могут быть нестабильными и сильно варьироваться.

5. ATE DR (59.974): метод двойной устойчивости (DR) комбинирует подходы из OLS и IPW, используя модели как исхода, так и propensity score.

Предпосылка: Либо модель исхода, либо модель propensity score правильно специфицирована.
Проблемы: Этот метод более устойчив к спецификационным ошибкам, но все же зависит от качества обеих моделей.

6. ATE dml_standard (61.5): двойное машинное обучение (DML), которое использует методы машинного обучения для оценки модели исхода и пропенсити скор.

Предпосылка: Обе модели (исхода и пропенсити скор) правильно специфицированы или имеют низкую ошибку предсказания.
Проблемы: Требует больших данных и хороших моделей машинного обучения. Если модели машинного обучения неправильно специфицированы или недостаточно мощные, оценки могут быть смещенными.

<u>Повышенная сложность:</u> включите дополнительный метод, не рассматривавшийся в курсе, и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

Мы оценили ATE с помощью Matching Estimator в Causal Model. Метод Matching используется для оценки среднего эффекта воздействия (ATE) путем сопоставления наблюдений из группы воздействия с наблюдениями из контрольной группы на основе схожих характеристик (ковариат).
Первым шагом является оценка propensity score - вероятность получения воздействия для $i-го$ наблюдения, вычисленная с использованием логистической регрессии или другого метода - для каждого наблюдения. Затем Для каждого наблюдения i из группы воздействия ($Last\_Ticket_i=1$) находим наблюдение j из контрольной группы ($Last\_Ticket_j=0$) , которое имеет наиболее схожий propensity score. Затем для каждой пары (i,j) оцениваем разность в целевой переменной. И, наконец, находим средний эффект воздействия (ATE), который оценивается как среднее значение разностей в целевой переменной для всех пар: $\text{ATE} = \frac{1}{n} \sum_{(i,j)} (\text{Ticket\_Price}_i - \text{Ticket\_Price}_j)$
Таким образом, оценка ATE_matching составила 62.259.
Теперь назовем некоторые преимущества и недостатки метода Matching Estimator по сравнению с другими методами.
Преимущества:

1. Matching Estimator не требует предположений о функциональной форме связи между воздействием и исходом, что делает его более гибким и применимым к различным ситуациям.
2. Метод учитывает неслучайный отбор воздействия, что позволяет уменьшить смещение оценок ATE, связанное с неслучайностью выборки.

Недостатки:

1. Качество оценки ATE с помощью Matching Estimator сильно зависит от точности сопоставления индивидуальных наблюдений в группах, получивших воздействие и контрольной группе. Неверное сопоставление может привести к смещению оценки.

2. Matching Estimator может быть ресурсоемким методом, особенно при больших выборках и многочисленных переменных. Необходимость в точном сопоставлении может потребовать больших объемов вычислительных ресурсов.

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.

Представим, что переменная $Population_i$ отсутствует в данных, из-за чего возникает эндогенность, поскольку эта переменная влияет и на цену авиабилета $Ticket\_Price_i$, и на количество свободных мест $Last\_Ticket_i$. Сначала мы воспользовались ДМО без инструментальной переменной, а потом с ней. Получили следующие результаты(см. Таблицу 16):

Таблица 16. Оценки LATE с использованием разных подходов.

	Оценка
LATE	43.46
LATE dml standard2	61.479
LATE dml iv	42.933

Различие между средним эффектом воздействия (ATE) и локальным средним эффектом воздействия (LATE) заключается в том, что ATE представляет собой средний эффект воздействия на всей выборке, включая как тех, кто получил воздействие, так и тех, кто нет, в то время как LATE оценивает эффект только на тех, кто реально получил воздействие. Локальный средний эффект воздействия (LATE) интерпретируется как изменение в цене авиабилета, вызванное изменением количества свободных мест на борту, наблюдаемое только среди тех рейсов, которые действительно получили воздействие. Таким образом, используя LATE мы можем оценить, как изменение количества свободных мест на борту влияет на цену авиабилета только среди тех пассажиров, которые реально купили билет на рейс с определенным количеством свободных мест, мы смотрим только на тех, кто действительно получил воздействие, и оцениваем эффект изменений цены авиабилета исключительно в этой подгруппе.

<u>Повышенная сложность:</u> воспользуйтесь также параметрической моделью, например, с помощью пакета switchSelection. Обсудите преимущества и недостатки такого подхода по сравнениию с двойным машинным обучением.

Для оценки LATE мы используем метод двухшагового МНК. Однако оценка получилась слишком большой и с отрицательным знаком (-1327) по сравнению с исходным LATE. Рассмотрим преимущества и недостатки 2SLS по сравнению с двойным машинным обучением (DML).

Преимущества IV2SLS:

1. Интерпретируемость: Результаты легко интерпретировать и объяснять.

2. Теоретическая основа: Хорошо исследованный метод с сильными теоретическими основаниями.
3. Простота: Относительно прост в реализации для стандартных задач.

Недостатки IV2SLS:

1. Чувствительность к выбору инструментов: Результаты сильно зависят от выбора инструментальных переменных. Плохие инструменты могут привести к смещению и неэффективности.
2. Линейные допущения: Предполагает линейность в отношении инструментальных переменных, что может быть ограничением в сложных моделях.
3. Меньшая гибкость: Не так гибок, как методы машинного обучения, для улавливания сложных нелинейных зависимостей.

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.

X-learner сочетает подходы T-learner и S-learner, используя их результаты для обучения двух моделей: одной на подверженной группе и одной на не подверженной. Затем он использует эти модели для получения взвешенного среднего среди оценок эффекта воздействия.

По результатам оценок, метод X-learner показывает некоторое улучшение (значения ближе к первоначальным CATE) по сравнению с другими методами, хотя не во всех случаях. Мотивация в его использовании состоит в том, чтобы использовать лучшую комбинацию подходов, полученную из T-learner и S-learner.

В бизнесе или государственных программах полученные оценки среднего эффекта воздействия могут использоваться для принятия более информированных решений о внедрении или корректировке программ и стратегий. Например, в случае авиакомпаний, оценка среднего эффекта воздействия может помочь определить оптимальные стратегии ценообразования в зависимости от доступности мест на борту.

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.

Сначала мы сравнили точность оценок по среднеквадратической ошибке(MSE0), для оценки прогнозной точности моделей использовали среднюю абсолютную ошибку (MAE), отметим, что более низкие значения MAE указывают на более точные

прогнозы, далее использовали псевдоисходы. На результатов мы видим, что OLS имеет наименьшее значение MSE0, что может указывать на то, что оно лучше всего соответствует истинным значениям CATE. (см. Таблицу 18) MAE показывает, насколько близки прогнозы модели к истинным значениям. Здесь мы видим, что OLS имеет наименьшее значение MAE, что указывает на более точные прогнозы этой модели. Величина псевдоисходов также может быть полезным показателем для оценки модели. Здесь IPW имеет наименьшее значение, что может указывать на лучшую способность модели учитывать различия в вероятностях получения воздействия. Таким образом, мы делаем выбор в пользу подхода OLS.

Таблица 18. Модели оценивания средних эффектов воздействия с использованием различных подходов.

	MSE0	**MAE**	**MSE1**
OLS	128.385	11.138	401677.272
T-learner	336.135	16.493	401128.899
S-learner	247.753	14.512	401059.867
IPW	11172.142	79.750	388052.371
X-learner	390.449	17.279	400832.973

8. Оцените средние эффекты воздействия и локальные средние эффекты воздействия используя худшие из обученных классификационных и регрессионных моделей. Сопоставьте результаты с теми, что были получены с помощью лучших моделей. Сделайте вывод об устойчивости результатов к качеству используемых методов машинного обучения.

При сравнении результатов подходов с использованием лучших и худших моделей машинного обучения регрессии и классификации можно сделать следующие выводы:

1. В целом, оценки, полученные с использованием лучших моделей, склонны быть ближе к истинным значениям, чем оценки, полученные с использованием худших моделей. (см. Таблицу 19) Это свидетельствует о том, что более точные модели обеспечивают лучшую аппроксимацию к истинным данным.

2. Различия в результатах:

- Результаты, полученные с использованием худших моделей, могут значительно отличаться от результатов, полученных с использованием лучших моделей. Это может быть вызвано недостаточной адаптацией худших моделей к данным, недостаточной точностью предсказаний или недообучением моделей.
- В некоторых случаях результаты худших моделей могут быть еще больше далеки от истинных значений, чем результаты лучших моделей. Это указывает на необходимость выбора подходящих моделей и методов машинного обучения для задачи.

Таким образом, хотя результаты могут отличаться в зависимости от выбранного метода машинного обучения, использование более точных и адаптированных моделей обычно обеспечивает более близкие к истинным значениям оценки эффектов воздействия.

Таблица 19. Оценки LATE с использованием лучших и худших моделей МО.

Модели	LATE	ДМО без IV	ДМО с IV
Лучшие	43.46	61.479	42.933
Худшие		62.012	60.017

СПИСОК ЛИТЕРАТУРЫ:

1. Diego Escobari Dynamic Pricing, Advance Sales, and Aggregate DemandLearning in Airlines // Journal of Industrial Economics, Forthcoming // 2011 Ссылка: https://www.researchgate.net/publication/256043023_Dynamic_Pricing_Advance_Sales_and_Aggregate_Demand_Learning_in_Airlines
2. Volodymyr Bilotkach Understanding Price Dispersion in the Airline Industry: Capacity Constraints and Consumer Heterogeneity // SSRN Electronic Journal 1 // 2005 Ссылка: https://www.researchgate.net/publication/228299229_Understanding_Price_Dispersion_in_the_Airline_Industry_Capacity_Constraints_and_Consumer_Heterogeneity