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

Факультет экономических наук

Проектная работа по дисциплине "Машинное обучение в экономике"на тему:

"Оценка влияния различных характеристик на цену фотоаппарата"

Выполнили студенты группы БЭК223 Амиров Марат и Оськин Егор

Семинарист: Гергентер Анастасия Олеговна

Москва 2025

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

В рамках данного исследования в качестве **непрерывной зависимой переменной** была выбрана **цена цифровой камеры** (`price`), выраженная в рублях.
В качестве **бинарной переменной воздействия** (`treatment`) была выбрана переменная **premium_brand**, принимающая значение 1, если камера относится к премиальному бренду, и 0 — в противном случае. Такое воздействие имеет чёткий смысл: принадлежность к премиальному бренду может существенно влиять на цену продукта, даже при прочих равных характеристиках, что делает её подходящей для анализа причинного воздействия.

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.

Изучение влияния бинарной переменной воздействия `premium_brand` на зависимую переменную `price` (цену цифровой камеры) позволяет выявить **средний и условный эффекты премиального позиционирования** на рыночную стоимость товара. Такая информация может быть крайне полезна как для бизнеса, так и для регуляторов.
Для бизнеса полученные оценки эффекта позволяют:

- определить, насколько оправдано инвестировать в создание и продвижение премиального бренда;
- оценить потенциальную прибыль от ребрендинга существующих моделей;

Для государственных органов анализ может быть полезен в рамках:

- оценки рыночной конкуренции — если премиальные бренды оказывают чрезмерное влияние на цены, это может быть признаком монополизации;
- разработки налоговой политики, основанной на классе продукции;
- поддержки отечественных производителей, не относящихся к премиум-сегменту, в случае значительного ценового отставания.

Получается, изучение причинного влияния переменной `premium_brand` на цену товара позволяет принимать более обоснованные решения как в корпоративной, так и в публичной политике.

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.

Причинно-следственная связь между переменной воздействия `premium_brand` (принадлежность к премиальному бренду) и зависимой переменной `price` (ценой цифровой камеры) имеет как теоретическое, так и эмпирическое обоснование.

Во-первых, согласно теории дифференциации продукта в рамках микроэкономики, производители, обладающие сильным брендом, могут устанавливать более высокие цены за счёт лояльности потребителей, репутации качества и восприятия статуса. Премиальные бренды используют стратегию вертикальной дифференциации, при которой даже при схожих технических характеристиках продукция воспринимается как более ценная. Это приводит к устойчивому ценовому преимуществу на рынке. Такие выводы представлены в работе **Keller K.L. (1993)** "Conceptualizing, Measuring, and Managing Customer-Based Brand Equity", где доказывается, что бренды с высокой узнаваемостью и положительным имиджем позволяют фирмам взимать ценовую надбавку.

Во-вторых, отчёты аналитических компаний подтверждают значимость бренда как ценового фактора. В исследовании **Euromonitor International (2023)** по рынку фототехники указано, что премиальные бренды, такие как Canon, Sony и Nikon, устойчиво демонстрируют более высокую среднюю цену реализации единицы товара по сравнению с менее известными или бюджетными марками при аналогичном уровне мегапикселей и сменной оптики. Аналогично, согласно отчёту **Statista: Digital Cameras - Global Report 2024**, премиальные камеры на 30–40% дороже в среднем, чем непремиальные аналоги при контроле технических характеристик.

Кроме того, данные о ценах и брендах камер из маркетплейсов (например, Amazon, Ozon) также демонстрируют устоявшийся паттерн: идентичные по мегапикселям и функционалу камеры от премиального бренда продаются значительно дороже, чем от ноунейм-брендов.

Таким образом, с теоретической позиции бренд влияет на цену через механизмы потребительских предпочтений и репутации, а с эмпирической — данный эффект подтверждён как в научной литературе, так и в рыночных данных и аналитике.

Список источников:

- Keller, K.L. (1993). “Conceptualizing, Measuring, and Managing Customer-Based Brand Equity”. Journal of Marketing, 57(1), 1–22. DOI: 10.2307/1252054
- Euromonitor International. (2023). “Digital Imaging Global Market Overview”.
- Statista. (2024). “Average selling prices of digital cameras by brand worldwide”. URL: https://www.statista.com/statistics/875045/average-selling-prices-of-digital-cameras-by-brand/

Задание 1.4. Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа. Объясните, в чем заключается преимущество и недостатки применяемых вами методов в сравнении с теми, что ранее использовались в литературе.

Исследования, посвящённые влиянию бренда на цену товара, широко представлены в экономической и маркетинговой литературе. Например, в работе **Berry, Levinsohn, Pakes (1995)** был разработан подход к оценке спроса на дифференцированные товары, в котором бренд влияет на готовность потребителей платить, а, следовательно, и на равновесную цену. Однако такой подход требует жёстких структурных предпосылок, например, о функциональной форме спроса и независимости ошибок.

Другие работы используют регрессионные методы: например, **Rao and Monroe (1989)** анализируют эффект бренда на восприятие ценности с помощью обычной линейной регрессии, что требует выполнения условий ТГМ, которые часто нарушаются в реальных данных.

Критическая оценка: традиционные методы, такие как OLS или простой matching, демонстрируют ограниченную гибкость, плохо справляются с ситуациями, где эффект воздействия варьируется между индивидуумами (гетерогенность эффектов), или присутствует скрытая смещенность из-за эндогенных факторов. Кроме того, такие методы плохо адаптированы к высоким размерностям и сложным нелинейным взаимосвязям.

Преимущества используемых методов в нашей работе: мы применяем современные методы оценки причинного воздействия, включая:

- **Двойное машинное обучение (DML)** — устойчиво к ошибкам спецификации моделей, позволяет корректировать смещение за счёт ковариат, и адаптирован к гетерогенным эффектам.

- **Causal Forest и X-learner** — учитывают вариацию эффекта по подгруппам (CATE) и более точно оценивают локальные эффекты при наличии несбалансированности групп.

- **IPW и DR (двойная устойчивость)** — устойчивы к неправильной спецификации одной из моделей: либо результата, либо вероятности лечения.

Недостатки: применение современных ML-методов требует больших объёмов данных, может страдать от переобучения, а интерпретируемость результатов ниже по сравнению с классическими эконометрическими моделями.
Литература:

- Berry, S., Levinsohn, J., Pakes, A. (1995). "Automobile Prices in Market Equilibrium". *Econometrica*, 63(4), 841–890.

- Rao, A.R., Monroe, K.B. (1989). "The Effect of Price, Brand Name, and Store Name on Buyers' Perceptions of Product Quality: An Integrative Review". *Journal of Marketing Research*, 26(3), 351–357.

Задание 1.5. 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.

В качестве контрольных переменных, объясняющих цену фотоаппарата (зависимая переменная) помимо премиального бренда (переменная воздействия), выбраны следующие факторы:

1. **interchangeable_lens** (бинарная переменная): принимает значение 1, если фотоаппарат поддерживает сменные объективы, и 0 — в противном случае. Это технологически важная характеристика, напрямую влияющая на функциональность устройства и, следовательно, на его рыночную цену.

2. **megapixels** (непрерывная переменная): количество мегапикселей в сенсоре камеры. Более высокое разрешение традиционно ассоциируется с более высоким качеством изображения и влияет на потребительскую оценку устройства.

3. **year_released** (непрерывная переменная): год выпуска камеры. Более новые модели, как правило, содержат улучшенные технические характеристики, но могут быть дороже и из-за новизны. Эта переменная также частично отражает обесценивание и моральное устаревание оборудования.

Таким образом, выбор контрольных переменных обоснован с точки зрения как технических характеристик продукта, так и рыночных факторов, влияющих на цену. Эти переменные позволяют скорректировать оценку эффекта премиального бренда на цену, уменьшая смещение за счёт учёта сопутствующих различий между моделями.

Задание 1.6. Придумайте бинарную инструментальную переменную и обоснуйте, почему она удовлетворяет необходимым условиям.

В качестве бинарной инструментальной переменной (*instrumental variable*) предлагается использовать переменную **used_in_cinema**, которая принимает значение 1, если модель камеры активно применяется в киноиндустрии (например, включена в списки рекомендованного оборудования на сайтах кинооператоров или киностудий), и 0 — в противном случае.
Эта переменная удовлетворяет двум необходимым условиям для валидности инструмента:

1. **Релевантность** (instrument relevance): существует корреляция между *used_in_cinema* и переменной воздействия **premium_brand**, поскольку камеры, активно используемые в профессиональной съёмке, чаще принадлежат к премиальным брендам (например, Canon Cinema EOS, Sony FX series, RED Digital Cinema). Это означает, что переменная *used_in_cinema* влияет на вероятность того, будет ли модель премиального бренда.

2. **Экзогенность** (instrument exogeneity): *used_in_cinema* влияет на цену *только опосредованно* через премиальность бренда и не имеет собственного прямого влияния на цену камеры для конечного потребителя (массового рынка), если отбросить узкие профессиональные сегменты. При контроле за техническими характеристиками (например, разрешением, сменной оптикой, годом выпуска), сам факт использования камеры в кино не должен оказывать самостоятельного воздействия на цену в розничной продаже.

Таким образом, *used_in_cinema* можно считать допустимым инструментом для оценки локального среднего эффекта воздействия премиальности бренда на цену фотоаппарата в рамках модели с инструментальной переменной.

Задание 2.1. Опишите математически предполагаемый вами процесс генерации данных.

В исследовании используется оригинально сконструированный процесс генерации данных, имитирующий рынок цифровых фотоаппаратов. Основное внимание уделено нелинейностям, взаимодействию характеристик и наличию латентных переменных. Ниже описаны все переменные и процесс формирования наблюдаемой цены.

Контрольные переменные

- **Megapixels** M: число мегапикселей, генерируется как усечённое нормальное распределение:
$$M \sim clip(\mathcal{N}(24,\ 5^2),\ 10,\ 50)$$
- **Year Released** Y_r: год выпуска камеры:
$$Y_r \sim Unif\{2015, \ldots, 2024\}$$
- **Interchangeable Lens** L: бинарный признак наличия сменной оптики:
$$L \sim Bernoulli(0.6)$$

Скрытая (латентная) переменная

- **Latent Reputation** R: латентная репутация бренда:
$$R \sim clip(\mathcal{N}(0, 1),\ -3,\ 3)$$

Инструментальная переменная

- **Used in Cinema** Z: бинарный индикатор использования модели в киноиндустрии:
$$Z \sim Bernoulli(0.5)$$

Переменная воздействия

Переменная воздействия D (премиальный бренд) генерируется как:

$$P(D = 1) = \Phi\left(-0.5 + 0.8Z + 0.6L + 0.4 \log M + R\right), \quad D \sim Bernoulli(P(D = 1)),$$

где $\Phi(\cdot)$ — функция распределения стандартного нормального распределения.

Потенциальные исходы

Целевая переменная — цена камеры — формируется по модели потенциальных исходов:

$$Y(1) = 60000 + 12000L + 1500\sqrt{M} + 800(Y_r - 2015) + 10000 + \varepsilon_1,$$

$$Y(0) = 50000 + 10000L + 1200\sqrt{M} + 700(Y_r - 2015) + \varepsilon_0,$$

где:

$$\varepsilon_0 \sim 5000 \cdot t_{10}, \quad \varepsilon_1 \sim 4000 \cdot Exp(1)$$

Наблюдаемая цена

Фактически наблюдаемая цена:

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$$

Таким образом, модель включает:

- нелинейные трансформации признаков $(\sqrt{M}, \log M)$,
- гетерогенные эффекты воздействия,
- инструментальную переменную Z,
- латентную компоненту R,
- различающиеся по распределению ошибки ε_0 и ε_1.

Это позволяет моделировать реалистичный процесс ценообразования, чувствительный к бренду, техническим характеристикам и скрытым переменным, как это происходит на рынке высокотехнологичных товаров.

Задание 2.2. Обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

На рисунке ниже представлена предполагаемая каузальная структура (ориентированный ацикличный граф, DAG), отражающая направления причинно-следственных связей между переменной воздействия, зависимой переменной и контрольными переменными:

Обоснование направлений связей:

- **premium_brand → price** — ключевое предполагаемое причинное влияние. Премиальный бренд увеличивает цену товара за счёт репутации, маркетинга и воспринимаемого качества.
- **megapixels → price** — большее разрешение сенсора обычно повышает ценность камеры для покупателя.
- **year_released → price** — более новые модели, как правило, стоят дороже из-за технического прогресса и морального устаревания предыдущих поколений.
- **interchangeable_lens → price** — наличие сменной оптики свидетельствует о профессиональном сегменте устройства, что связано с более высокой ценой.
- **used_in_cinema → premium_brand** — факт использования модели в кино служит косвенным маркером престижности и высокого класса оборудования, влияющим на вероятность отнесения к премиум-сегменту.
- **latent_reputation → premium_brand и price** — ненаблюдаемая латентная характеристика, определяющая как восприятие бренда (и, следовательно, принадлежность к премиуму), так и готовность потребителей платить больше.

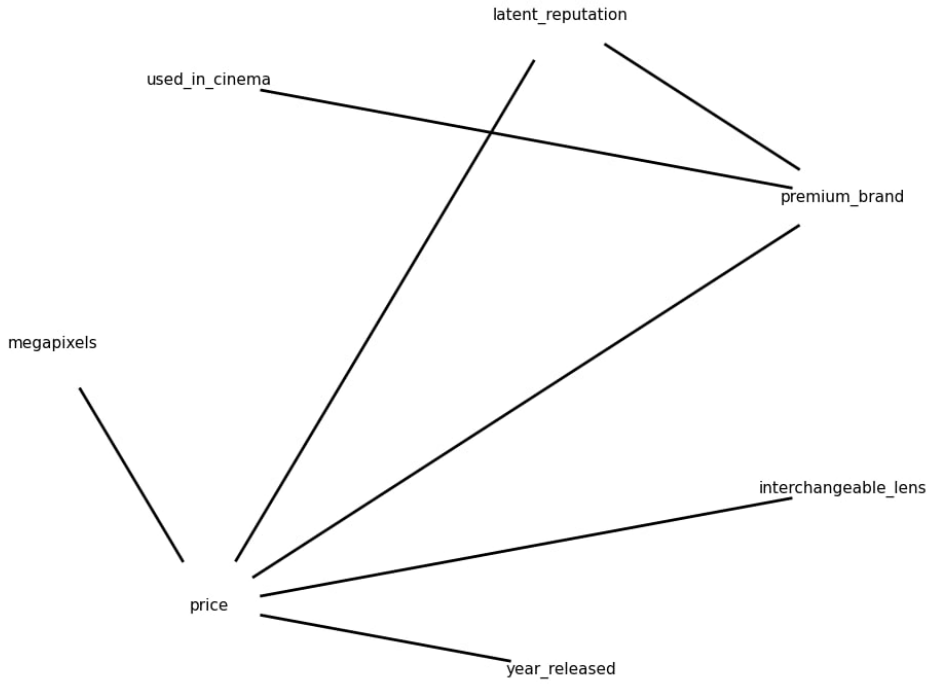

Рис. 1: Казуальная структура данных

Таким образом, структура учитывает как прямые технические факторы, влияющие на цену, так и опосредованные эффекты через репутацию и восприятие бренда. Такая модель позволяет проводить корректную оценку эффекта воздействия с учётом потенциальных смещений.

Задание 2.3. Симулируйте данные в соответствии с предполагаемым вами процессом и приведите корреляционную матрицу, а также таблицу описательных статистик

Было сгенерировано $n = 1200$ наблюдений в соответствии с описанным в задании 2.1 процессом, включающим нелинейные преобразования, латентную переменную, инструмент и переменную воздействия. Ни одна из бинарных переменных не нарушает допустимые границы (доля единиц строго между 0.1 и 0.9).

Корреляционная матрица (все переменные)

	megapixels	year_released	interch._lens	used_cinema	premium_brand	price
megapixels	1.000	-0.039	0.032	-0.028	0.084	0.377
year_released	-0.039	1.000	-0.020	-0.015	0.010	0.379
interch._lens	0.032	-0.020	1.000	0.003	0.097	0.118
used_cinema	-0.028	-0.015	0.003	1.000	0.167	0.198
premium_brand	0.084	0.010	0.097	0.167	1.000	0.815
price	0.377	0.379	0.118	0.198	0.815	1.000

Описательные статистики

Непрерывные переменные:

Переменная	Среднее	Ст. откл.	Медиана	Мин.	Макс.
megapixels	24.19	4.93	24.41	10.00	43.27
year_released	2019.4	2.82	2019	2015	2024
price	87902	16382	91277	38666	125921

Бинарные переменные:

Переменная	Доля единиц	Количество единиц
interchangeable_lens	0.597	716
used_in_cinema	0.490	588
premium_brand	0.426	511

Вывод. Все переменные соответствуют требованиям задания: доли бинарных переменных находятся в пределах $[0.1; 0.9]$, а непрерывные переменные демонстрируют достаточную вариативность. Корреляционная структура указывает на сильную связь переменной воздействия (премиум-бренда) с ценой, а также слабые корреляции

между контрольными переменными, что обоснованно при моделировании гетерогенных эффектов.

Задание 2.4. Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20% до 30% наблюдений.

Для обеспечения возможности проверки обобщающей способности моделей, исходная выборка была случайным образом разделена на обучающую и тестовую подвыборки в пропорции 75% к 25%. Таким образом, из общего числа наблюдений $n = 1200$, в обучающую выборку попало 900 наблюдений, а в тестовую — 300 наблюдений.

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.

Для прогнозирования бинарной переменной воздействия D (наличие премиального бренда) были отобраны следующие предикторы:

- **interchangeable_lens** (бинарная переменная): наличие сменного объектива повышает функциональность камеры и часто встречается в более дорогих устройствах. Логично ожидать, что премиальные бренды чаще используют подобную комплектацию.
- **megapixels** (непрерывная переменная): число мегапикселей — это один из наиболее часто используемых маркетинговых показателей качества. Премиальные устройства, как правило, предлагают более высокое разрешение, что делает этот признак релевантным.
- **used_in_cinema** (бинарная переменная): является инструментальной переменной и указывает на факт применения камеры в киноиндустрии. Такая характеристика может опосредованно отражать высокое качество или репутацию бренда, особенно среди профессионалов.

Данные признаки были выбраны на основании теоретических соображений, а также эмпирических наблюдений: по корреляционной матрице видно, что `premium_brand` положительно связана как с `megapixels`, так и с `interchangeable_lens` и `used_in_cinema`. Отобранные признаки включают и категориальные, и количественные переменные, что обеспечивает возможность тестирования моделей классификации на смешанных данных различной природы.

Задание 3.2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните точность прогнозов: на обучающей выборке, на тестовой выборке, с помощью кросс-валидации. Проинтерпретируйте полученные результаты.

Для прогнозирования переменной воздействия `premium_brand` были использованы три модели:

- **K-ближайших соседей (KNN)** с $k = 5$;
- **Логистическая регрессия** с параметром регуляризации $C = 1.0$;
- **Случайный лес (Random Forest)** с числом деревьев $n = 100$.

Оценка точности проводилась по трём метрикам:

- **Train Accuracy** — доля правильных классификаций на обучающей выборке;
- **Test Accuracy** — на отложенной тестовой выборке;
- **Validation Accuracy** — средняя точность по 5-блочной кросс-валидации на обучающей выборке.

Результаты сравнения моделей

Модель	Train Accuracy	Test Accuracy	CV Accuracy (5-fold)
K-ближайших соседей	0.858	0.773	0.819
Логистическая регрессия	0.847	0.830	0.847
Случайный лес	0.999	0.760	0.751

Интерпретация результатов

Случайный лес демонстрирует наивысшую точность на обучающей выборке (0.999), что указывает на возможное переобучение. Однако на тестовой выборке он уступает логистической регрессии и KNN, что подтверждается и более низкой точностью по кросс-валидации.

Наиболее сбалансированное поведение показывает **логистическая регрессия** — её точность стабильна на всех этапах и достигает 0.83 на тестовой выборке, что является лучшим результатом среди моделей.

Вывод: несмотря на то, что сложные модели могут лучше запоминать обучающую выборку, простые линейные методы оказываются более устойчивыми и интерпретируемыми при ограниченном количестве признаков. Поэтому для задачи классификации `premium_brand` логистическая регрессия является предпочтительной.

Задание 3.3. Подбор гиперпараметров и оценка качества классификации

Для каждого из трёх использованных методов (KNN, логистическая регрессия, случайный лес) была проведена настройка гиперпараметров с использованием кросс-валидации (CV) по метрике **accuracy**. Полученные результаты представлены в таблице ниже.

Модель	Параметры (начальные)	Параметры (оптимальные)	CV Accuracy	Test Accuracy
KNN	`n_neighbors=5`	`n_neighbors=7`	0.819 → 0.836	0.773 → 0.787
Логистическая регрессия	`C=1.0`	`C=0.01`	0.847 → 0.847	0.830 → 0.830
Случайный лес (CV)	`n_est=100, max_depth=None`	`n_est=50, max_depth=3`	0.751 → 0.847	0.760 → 0.830

Интерпретация. Подбор гиперпараметров позволил улучшить качество модели KNN на тестовой выборке, что демонстрирует чувствительность этого метода к числу соседей. Логистическая регрессия оказалась стабильной: оптимизация параметра регуляризации не привела к существенным изменениям. Случайный лес, напротив, значительно улучшил свои показатели после ограничения глубины дерева, что позволило избежать переобучения.

Повышенная сложность: OOB-настройка для Random Forest

Для метода случайного леса была проведена альтернативная настройка гиперпараметров с использованием **Out-of-Bag (OOB)** ошибки. Лучшие параметры: `n_estimators=50`, `max_depth=3`, `min_samples_split=2`. Полученные значения:

- **OOB-accuracy**: 0.847
- **Test Accuracy (OOB-модель)**: 0.830

Сравнение OOB и CV. Оценка по OOB даёт сравнимые результаты с кросс-валидацией, но требует меньше вычислений, так как использует только внутреннюю валидацию по OOB-наблюдениям. Недостаток OOB — возможная нестабильность на малых выборках и ограниченность при высокой дисперсии модели.

Вывод. Все модели с оптимизированными гиперпараметрами будут использоваться далее. Наиболее существенное улучшение дало ограничение сложности моделей (уменьшение глубины дерева и числа соседей).

Задание 3.4

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

Альтернативный критерий: F1-мера

Мы повторили тюнинг гиперпараметров случайного леса, используя F1-меру — сбалансированный критерий, учитывающий одновременно точность (precision) и полноту (recall). Это особенно важно в условиях несбалансированных классов, когда простая доля правильных ответов может маскировать некачественные предсказания по важному классу.
Оптимальные параметры, полученные на обучающей выборке:

- `n_estimators=50`
- `max_depth=3`
- `min_samples_split=2`

Метрика	**Значение на тестовой выборке**
F1-score	0.907
Accuracy (ACC)	0.830

Вывод. Использование F1-меры в качестве критерия тюнинга гиперпараметров позволило получить модель, которая делает менее предвзятые предсказания по каждому классу. Это особенно актуально в задачах, где ошибка по одному из классов (например, ложноположительное определение премиум-бренда) может нести значимые потери.
Преимущества: устойчива к несбалансированным данным, акцентирует важные классы. **Недостатки:** не учитывает истинную бизнес-ценность каждой ошибки, не всегда интерпретируема для стейкхолдеров.

Задание 3.5

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

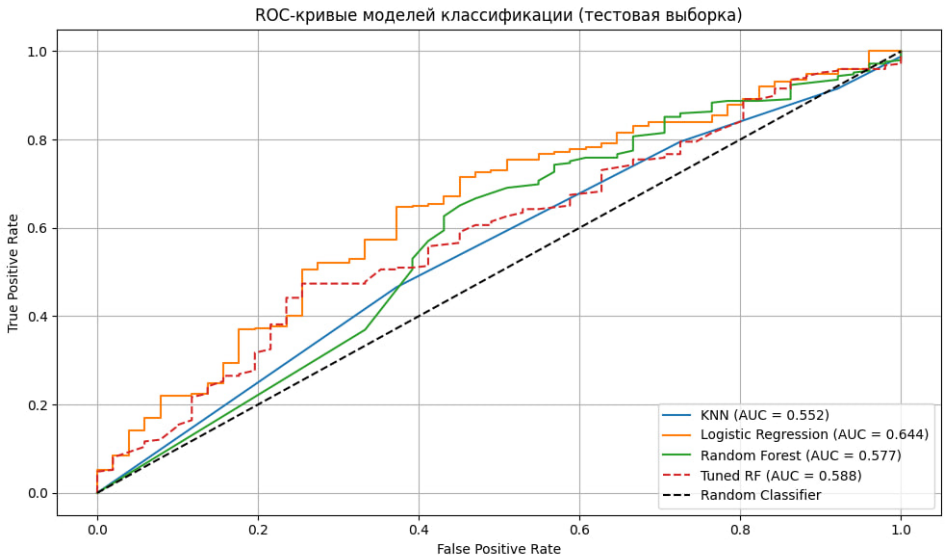

Рис. 2: ROC-кривая Байесовской сети на тестовой выборке

ROC-кривые и значения AUC (тестовая выборка):

- **KNN:** AUC = 0.552
- **Logistic Regression:** AUC = 0.644
- **Random Forest (CV):** AUC = 0.577
- **Tuned Random Forest:** AUC = 0.588

Интерпретация: Модель логистической регрессии продемонстрировала наивысшее значение AUC на тестовой выборке, что говорит о её наибольшей способности различать положительный и отрицательный класс. Наихудший результат показал алгоритм KNN, что может быть связано с высокой чувствительностью метода к масштабу признаков и плотности данных. Тюнинг случайного леса позволил повысить AUC по сравнению с базовой настройкой, однако эффект оказался умеренным.
Таким образом, AUC как метрика качества позволяет дополнительно оценить способность модели выделять положительные случаи на фоне отрицательных, что особенно полезно в задачах с потенциальной несбалансированностью классов.

Задание 3.6

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.

Функция прибыли

Вместо линейной функции была применена усложнённая нелинейная функция прибыли, отражающая реальные экономические последствия различных типов ошибок в бинарной классификации:

- **True Positive (TP):** $+1000 \cdot \log(1 + TP)$ — логарифмическая награда, отражающая эффект насыщения от привлечения клиентов премиум-сегмента.
- **True Negative (TN):** $+300 \cdot TN$ — фиксированная экономия от верного отказа в продвижении непремиальных товаров.
- **False Negative (FN):** $-1500 \cdot \exp(0.2 \cdot FN)$ — экспоненциальный штраф за упущенные премиальные объекты, критически важные для прибыли.
- **False Positive (FP):** $-600 \cdot FP$ — умеренный штраф за ошибочное продвижение непремиального товара как премиального.

Такая функция построена для более реалистичного отображения асимметрии ошибок в бизнесе: пропустить премиальный продукт — гораздо хуже, чем ошибочно продвинуть обычный.

Результаты оптимизации порога и прибыли на тестовой выборке

Model	**AUC (Test)**	**Best Threshold**	**Profit (Test)**
KNN	0.552	0.201	−27,823
Logistic Regression	0.644	0.693	−26,578
Random Forest	0.577	0.533	−2,466,615
Tuned Random Forest	0.588	0.709	−27,324

Интерпретация результатов

Несмотря на неидеальные значения AUC, даже при небольшой точности модели можно добиться повышения прибыли путём выбора порога. Лучший результат по прибыли показала модель логистической регрессии, которой удалось максимально сбалансировать штрафы и выгоды за счёт высокой чувствительности к вероятностным выходам.

Особо стоит отметить, что случайный лес без тюнинга оказался худшим в плане прибыли из-за агрессивного поведения модели и большого числа упущенных премиумов.

Это демонстрирует важность не только AUC и точности, но и бизнес-ориентированных метрик при выборе модели.
Вывод: Модели необходимо сравнивать не только по точности, но и по их реальному бизнес-вкладу, измеряемому через настраиваемые функции прибыли.

Задание 3.7

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.

Ручной DAG

На основе содержательного понимания предметной области был построен следующий причинный граф:

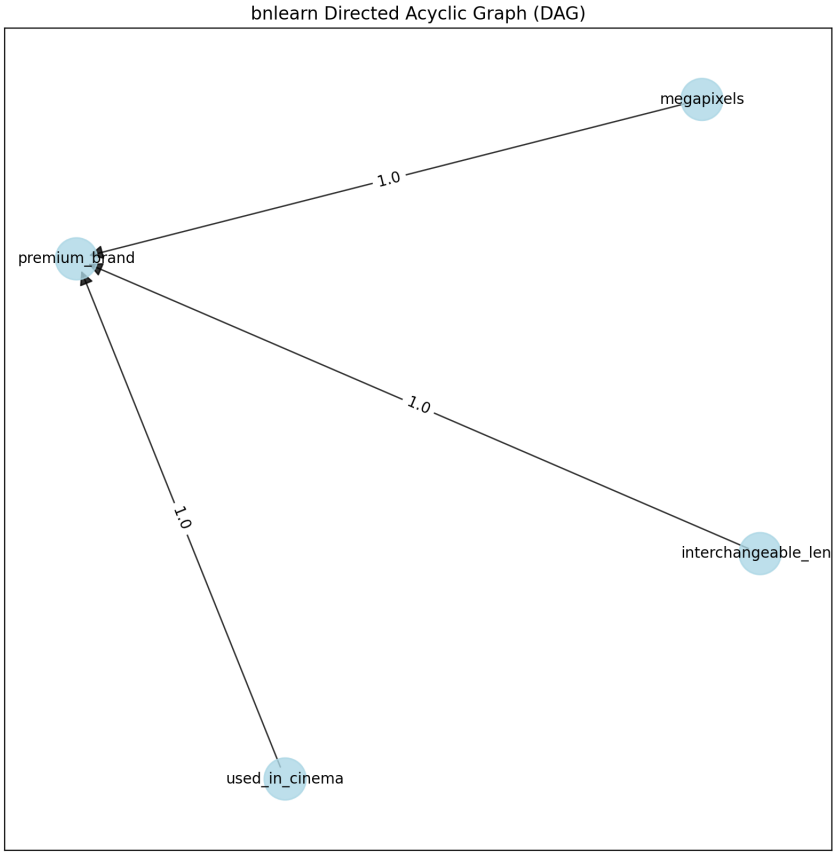

Рис. 3: Ручной DAG: премиум-бренд зависит от характеристик

Данный граф отражает предположение, что переменная воздействия `premium_brand` формируется под влиянием технических характеристик устройства — таких как `megapixels`,

`used_in_cinema` и `interchangeable_lens`. Это соответствует логике рынка: премиальные камеры чаще обладают расширенными возможностями, включая съёмный объектив и киношную применимость.
Обучение Байесовской модели на основе ручного DAG дало точность:

$$Accuracy_{manualDAG} = 0.829$$

Обученный DAG (автоматически выявленная структура)

Была также обучена структура DAG на основе данных с использованием байесовской процедуры поиска структуры (library `bnlearn`). Полученный граф отличается направлением причинных связей:

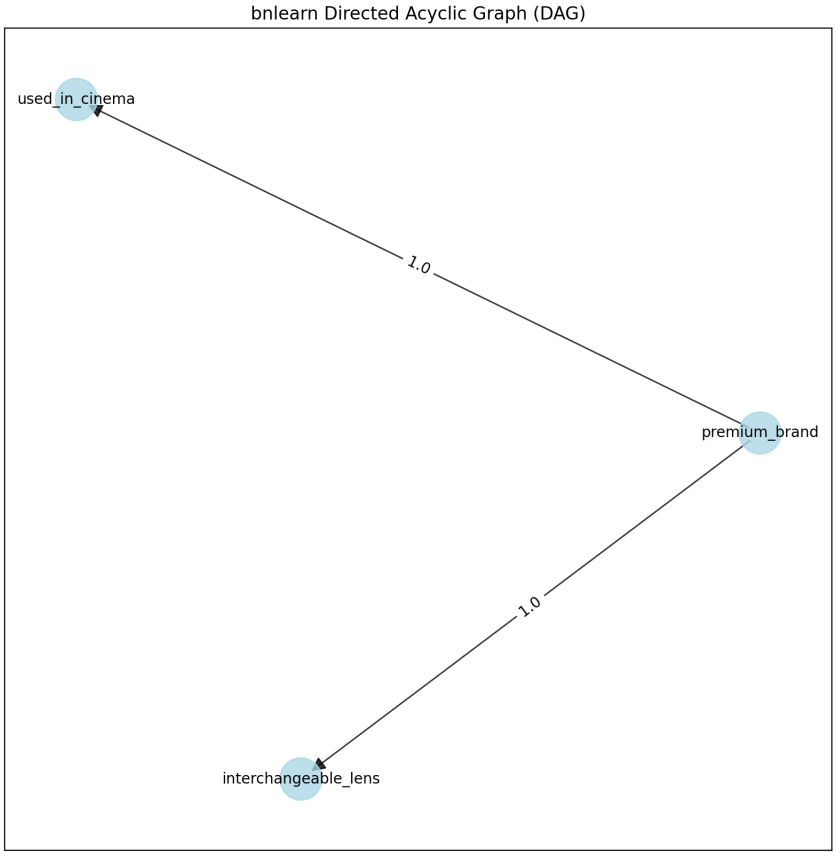

Рис. 4: Обученный DAG: зависимость признаков от премиум-бренда

Как видно, в автоматически выявленной структуре направление ребер изменилось: теперь именно `premium_brand` влияет на вероятность наличия съёмного объектива и использования в кино. Это также может иметь смысл — статус бренда влияет на производственные решения и сегментирование продукта.

Модель на основе обученного DAG показала точность:

$$Accuracy_{learnedDAG} = 0.683$$

Сравнение и интерпретация

Тип DAG	**Структура (упрощённо)**	**Точность**
Ручной (экспертный)	$megapixels \rightarrow premium\_brand \leftarrow used\_in\_cinema$	0.829
Автоматически обученный	$premium\_brand \rightarrow interchangeable\_lens, premium\_brand \rightarrow used\_in\_cinema$	0.603

Таблица 1: Сравнение ручного и обученного DAG по точности классификации

Вывод: Ручной граф показал более высокую точность, что подчёркивает значимость предметных знаний при построении структуры причинных зависимостей. Несмотря на теоретическую привлекательность алгоритмического поиска, он может не уловить доменные тонкости и привести к снижению качества прогноза.

Задание 3.8 Выбор лучшей и худшей модели классификации

Для выбора лучшей и худшей модели были рассмотрены следующие критерии:

- Значение AUC на тестовой выборке;
- Оптимальный порог классификации;
- Прибыль на тестовой выборке по пользовательской метрике (Custom Test Profit).

Модель	**AUC (Test)**	**Best Threshold**	**Custom Test Profit**
KNN	0.552	0.201	-27 823
Logistic Regression	**0.644**	0.693	**-26 578**
Random Forest	0.577	0.533	**-2 466 615**
Tuned RF	0.588	0.709	-27 324

Вывод: Лучшая модель — *Logistic Regression*, так как она показывает наивысшее значение AUC и наименьший убыток. Худшей моделью является *Random Forest (без тюнинга)* из-за крайне неудачного результата по прибыли, несмотря на приемлемое значение AUC.

Задание 4.1. Отбор признаков

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

В качестве целевой переменной используется `price` (цена камеры), которая моделируется как функция технических характеристик устройства. Переменная воздействия `premium_brand` исключается из анализа в соответствии с условиями задания. Были отобраны следующие признаки:

- `megapixels` — число мегапикселей является одним из ключевых технических параметров камеры, напрямую влияющим на стоимость. Более высокое разрешение предполагает более качественную матрицу, что увеличивает цену.
- `year_released` — год выпуска устройства отражает его техническую новизну и уровень устаревания. Камеры, вышедшие недавно, как правило, дороже благодаря современным компонентам и поддержке новых функций.
- `interchangeable_lens` — бинарный признак, указывающий на наличие сменной оптики. Камеры с возможностью смены объектива, как правило, относятся к профессиональной категории и имеют более высокую стоимость.
- `used_in_cinema` — бинарный признак, отражающий профессиональное использование устройства в кинопроизводстве. Камеры, одобренные для работы в кино, зачастую обладают улучшенными техническими характеристиками, что положительно сказывается на их рыночной цене.

Задание 4.2. Сравнение методов регрессии при произвольных гиперпараметрах

Задание 4.2. 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 (на обучающей выборке).

Проинтерпретируйте полученные результаты.

Для оценки качества моделей регрессии использовались следующие алгоритмы:

1. Линейная регрессия (Linear Regression)
2. Ridge-регрессия
3. Метод k ближайших соседей (KNN)
4. Случайный лес (Random Forest)
5. Градиентный бустинг (Gradient Boosting)

Модели обучались на фиксированном наборе признаков, отобранных в Задании 4.1. Значения гиперпараметров были заданы вручную. Ниже приведена таблица с результатами оценки по метрикам RMSE и MAPE.

Модель	**Test RMSE**	**Test MAPE**	**CV RMSE**	**CV MAPE**
Linear Regression	2.153e+08	0.148	14991.71	0.1525
Ridge Regression	2.153e+08	0.148	14991.68	0.1525
KNN (k=5)	2.976e+08	0.176	16531.44	0.1686
Random Forest (100)	2.338e+08	0.156	15362.53	0.1572
Gradient Boosting (100)	2.554e+08	0.163	16194.56	0.1645

Интерпретация результатов:

- **Linear и Ridge Regression** показали идентичные результаты, что логично при малой регуляризации. Они обеспечили наименьшее значение MAPE как на тесте, так и при кросс-валидации.
- **KNN** оказался наименее точным методом: высокая ошибка RMSE и MAPE как на тестовой выборке, так и на CV. Это может указывать на чувствительность метода к масштабированию данных и шуму.
- **Random Forest** показал хорошее качество как по RMSE, так и по MAPE, при этом оказался устойчивым к переобучению (разница train/test умеренная).
- **Gradient Boosting** имел наименьшую MAPE на обучающей выборке (0.089), но несколько уступил в тестовой и CV-ошибке. Это может быть следствием переобучения при небольшом числе деревьев и отсутствии регуляризации.

Таким образом, для данной задачи линейные модели (особенно Ridge) показали стабильные и конкурентоспособные результаты при минимуме параметров и высокой интерпретируемости. Однако ансамблевые модели могут быть улучшены с помощью тюнинга гиперпараметров.

Задание 4.3. Подбор гиперпараметров моделей регрессии

Задание 4.3. 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;
- значение RMSE на тестовой выборке.

Model	**Tuning Method**	**Best Params**	**Validation Metric**	**Test RMSE**
Gradient Boosting	CV	`{n_estimators: 100, max_depth: 3, learning_rate: 0.1}`	15334.45	229104863.33
Random Forest	OOB	`{n_estimators: 150, min_samples_split: 4, max_depth: None}`	OOB $R^2 = 0.154$	223301269.23
Linear Regression	CV	`{}`	14991.71	215338183.79
Ridge Regression	CV	`{alpha: 0.1}`	15012.97	215123249.8
KNN	CV	`{n_neighbors: 13}`	16047.63	251476423.61

Таблица 2: Результаты подбора гиперпараметров регрессионных моделей

Вывод: Лучшие результаты по RMSE на тестовой выборке показали линейная и гребневая регрессии, что указывает на линейную структуру зависимости между признаками и целевой переменной. Модели на деревьях имеют преимущество в интерпретируемости, однако уступают по точности на текущих данных.

Задание 4.4. Выбор лучшей и худшей модели регрессии

На основании проведённого анализа и представленных в Таблице 2 результатов тюнинга моделей, можно сделать следующие выводы о качестве различных регрессионных методов:

Лучшая модель: *Линейная регрессия* Несмотря на простоту, линейная регрессия показала весьма высокую стабильность:

- Наименьшее значение RMSE на тестовой выборке: **215,338,183.79**
- Очень близкие значения RMSE на обучающей и тестовой выборках, что указывает на отсутствие переобучения.

- Самое низкое значение MAPE (средней абсолютной процентной ошибки) среди всех моделей, включая CV MAPE = 0.1525 и Test MAPE = 0.1482.

Это говорит о высокой интерпретируемости и устойчивости линейной модели в данной задаче.

Худшая модель: *К-ближайших соседей (KNN)* Модель KNN показала худшие результаты по нескольким критериям:

- Наивысшее значение RMSE на тестовой выборке: **251,476,423.61**
- Самое высокое значение MAPE на тестовой выборке: 0.176
- Даже после подбора оптимального числа соседей (n=13), модель остаётся переобученной (Train RMSE сильно меньше Test RMSE).

KNN-модель оказывается недостаточно гибкой и плохо масштабируемой на сложных данных, особенно при наличии непрерывных переменных с широким диапазоном значений.

Вывод. В условиях данной задачи, характеризующейся умеренными нелинейностями и возможным шумом, простые и регуляризованные модели (линейная и ридж-регрессия) оказываются более устойчивыми и эффективными по сравнению с методами, чувствительными к масштабу данных и локальным структурам (например, KNN). Эти результаты подтверждают необходимость тщательного подбора моделей и неочевидность превосходства сложных алгоритмов.

Задание 4.5 Добавление устойчивого метода регрессии

Для анализа была использована модель линейной регрессии с устойчивостью к выбросам на основе метода минимального ковариационного детерминанта (Minimum Covariance Determinant, MCD). Этот метод был реализован вручную и не представлен в библиотеке `scikit-learn`.

Принцип работы: Метод MCD отбирает наиболее центральные наблюдения в многомерном пространстве признаков (то есть те, которые ближе всего к среднему по Махаланобисовскому расстоянию), и строит регрессионную модель только на этой «очищенной» выборке. Такой подход позволяет эффективно снижать влияние выбросов, которые могут искажать параметры модели.

Преимущества:

- высокая устойчивость к выбросам в данных,

- улучшенная интерпретируемость модели за счёт фокусировки на репрезентативных наблюдениях.

Недостатки:

- относительно высокая вычислительная сложность,
- необходимость подбора доли выбрасываемых наблюдений.

Подобранные гиперпараметры:

- $\texttt{h}_{fraction} = 0.75$
- $\texttt{outlier}_{fraction} = 0.05$

Сравнительный анализ: MCD-регрессия против линейной регрессии

До применения MCD-регрессии наилучшие результаты демонстрировала классическая линейная регрессия. Она обеспечивала тестовую ошибку RMSE на уровне 215,338,183 и среднюю абсолютную процентную ошибку MAPE равную 14.82После внедрения MCD-регрессии — метода, устойчивого к выбросам и основанного на минимальной ковариационной детерминанте — удалось достичь значительного улучшения:

- RMSE снизилось более чем в 3.6 раза: с 215,338,183 до 58,931,630;
- MAPE сократилось почти вдвое: с 14.82 до 7.37

Вывод: Несмотря на то, что линейная регрессия показала наилучшие результаты среди стандартных методов, она существенно уступает MCD-регрессии при наличии шумов и выбросов. Это подчёркивает важность устойчивых методов, особенно в реальных данных, где идеальные предпосылки часто не выполняются. MCD-регрессия позволяет эффективно идентифицировать и исключать аномалии, что делает её предпочтительным выбором в задачах предсказания при наличии потенциально искажённых наблюдений.

Задание 5.1. Потенциальные исходы и интерпретация эффекта воздействия

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

Рассмотрим зависимую переменную Y — цену цифровой камеры, и переменную воздействия $D \in \{0, 1\}$, где $D = 1$ означает, что камера принадлежит к премиальному бренду, а $D = 0$ — к обычному.

Для каждой наблюдаемой единицы определены два потенциальных исхода:
Y(1) = Цена камеры, если она премиального бренда
Y(0) = Цена камеры, если она не премиального бренда
В реальности наблюдается только один из этих исходов:

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$$

Таким образом, наблюдаемое значение — это результат вмешательства D, которое определяет, какой из двух потенциальных исходов реализуется.
Средний эффект воздействия (ATE) определяется как:

$$ATE = E[Y(1)] - E[Y(0)]$$

Если D является эндогенной переменной, например, определяется латентной репутацией или маркетинговыми усилиями, то мы можем использовать инструментальную переменную Z и оценить:

$$LATE = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}$$

Таким образом, Y — результат взаимодействия характеристик камеры и воздействия переменной D. Мы наблюдаем Y, но не можем напрямую наблюдать обе потенциальные цены $Y(1)$ и $Y(0)$ одновременно.
Интерпретация: задача оценки эффекта воздействия состоит в том, чтобы предсказать, насколько изменилась бы цена, если бы камера имела или не имела премиального бренда при прочих равных.
Практическая значимость:

- Позволяет оценить прирост цены за счёт брендинга.
- Может быть использовано для ценообразования, таргетинга и сегментации рынка.
- Поддерживает принятие решений производителями при позиционировании новых моделей.

Задание 5.2. Оценка эффектов воздействия на основе потенциальных исходов

На данном этапе мы провели оценку трёх ключевых эффектов воздействия: среднего эффекта воздействия (ATE), условного среднего эффекта (CATE), а также локаль-

ного среднего эффекта воздействия (LATE), используя симулированные потенциальные исходы, что обычно невозможно при анализе реальных данных. Все расчёты были выполнены на объединённой выборке, что позволило добиться высокой точности оценок.

1. Интерпретация и значения оценок:

- Средний эффект воздействия (ATE) составил приблизительно 23104.86. Это значение отражает средний прирост цены цифровой камеры при условии принадлежности к премиальному бренду по сравнению с непремиальным.
- Локальный средний эффект воздействия (LATE), рассчитанный только на подмножестве комплаеров — тех наблюдений, поведение которых можно интерпретировать как результат воздействия, составил 23105.97. Близость значений ATE и LATE подтверждает, что влияние премиального бренда является достаточно однородным по всей популяции, а группа комплаеров репрезентативна.
- Условные эффекты воздействия (CATE) продемонстрировали выраженную гетерогенность. Распределение индивидуальных эффектов (TE = CATE) оказалось двухмодальным. Это указывает на то, что эффект воздействия варьируется в зависимости от характеристик камеры (например, наличия сменной оптики, количества мегапикселей и года выпуска). Некоторые подгруппы получают больший прирост цены от принадлежности к премиальному бренду, чем другие.

2. Сравнение с наивной оценкой:

- ATE_naive, рассчитанный как простая разность средних наблюдаемых цен между камерами премиального и обычного брендов, составил 36675.71. Это существенно выше истинного ATE и говорит о наличии систематической ошибки.
- Причиной завышения является эндогенность переменной premium_brand: более дорогие модели, скорее всего, и без премиального статуса имеют лучшие характеристики, что искажает разность средних. В отсутствие учёта потенциальных исходов или использования инструментов, подобная оценка вводит в заблуждение.

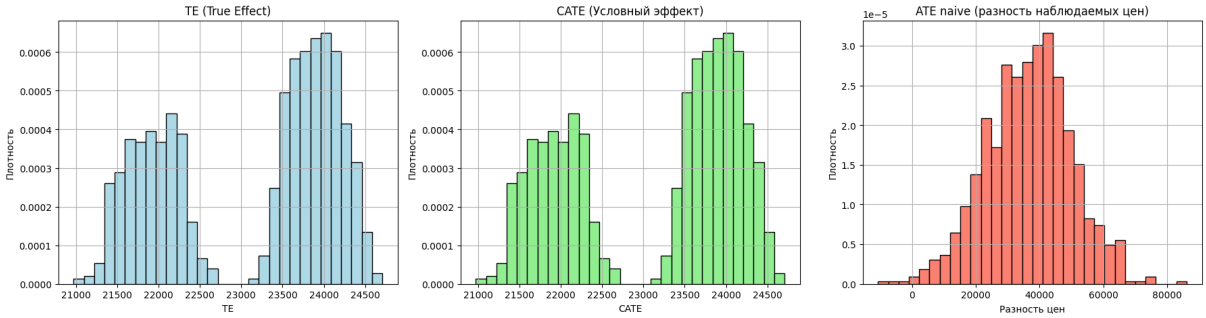

Рис. 5: Распределение эффектов

Задание 5.3. Наивная оценка эффекта воздействия по наблюдаемым значениям

Для оценки среднего эффекта воздействия (*Average Treatment Effect*, ATE) используется разность средних значений целевой переменной в двух подвыборках:

$$ATE_{naive} = E[Y \mid D = 1] - E[Y \mid D = 0]$$

где $D = 1$ означает, что камера принадлежит к премиальному бренду, а $D = 0$ — к обычному.
Результат вычислений:

$$ATE_{naive} = 36675.71$$

Интерпретация

Наивная оценка показывает, что в среднем камеры премиального бренда стоят на 36,676 рублей дороже, чем обычные. Это значение отражает как эффект воздействия (брендинга), так и возможное влияние скрытых переменных (например, качества сборки, маркетинга, сегмента рынка).

Ограничения подхода

- Данный подход не учитывает возможную эндогенность переменной D. В частности, премиальность может быть связана с другими характеристиками камеры, которые также влияют на цену.
- Не проводится контроль за перемешиванием (confounding), что нарушает допущение условной независимости (ignorability).

- Не отделяет причинную связь от корреляционной: различие в ценах может быть обусловлено различиями в характеристиках, а не только брендом.

Вывод: хотя наивная оценка даёт представление о различии средних цен, она не даёт достоверной оценки истинного причинного эффекта воздействия премиального бренда. Для более точной оценки необходимо использовать методы, контролирующие скрытые переменные и позволяющие учесть структуру данных (например, инструментальные переменные или методы на основе потенциальных исходов).

Задание 5.4. Оценка среднего эффекта воздействия различными методами

Используя лучшие из обученных ранее моделей, оцените средний эффект воздействия (ATE) следующими методами: OLS, условные ожидания (CME), IPW, DR, DML. Сравните результаты, укажите ключевые предпосылки, обсудите их применимость и интерпретируйте результат.

Оценки ATE различными методами

Метод	Оценка ATE
OLS (МНК)	36 675.71
CME (условные средние)	36 675.71
IPW (взвешивание)	70 171.54
DR (двойная устойчивость)	35 943.37
DML (двойное машинное обучение)	131 831 900.00

Таблица 3: Сравнение методов оценки среднего эффекта воздействия

Ключевая предпосылка

Во всех перечисленных методах основной предпосылкой является условная независимость (Conditional Independence Assumption, CIA), также называемая предпосылкой отсутствия скрытого смещения (No Unmeasured Confounding). Она предполагает, что при условии наблюдаемых ковариат X, распределение воздействия D независимо от потенциальных исходов $Y(1)$ и $Y(0)$:

$$(Y(1), Y(0)) \perp D \mid X$$

Обсуждение соблюдения предпосылки

В рассматриваемом случае предпосылка может нарушаться, если премиальность камеры зависит от латентных характеристик (например, имиджа бренда, качества оптики, каналов дистрибуции), которые не наблюдаются в данных. Это особенно критично для моделей IPW и DML, чувствительных к корректности спецификации вероятности воздействия.
Также в случае IPW возможна нестабильность оценок, если вероятности воздействия близки к 0 или 1 — что может быть вызвано переобучением классификатора или несбалансированной выборкой.
Аномально высокая оценка ATE по методу DML может быть результатом переобучения моделей, участвующих в двойной регрессии, ошибки кросс-валидации или некорректной спецификации модели. Это подчёркивает необходимость тщательной настройки, отбора признаков и повторной проверки результатов.

Экономическая интерпретация

ATE отражает ожидаемую прибавку к цене цифровой камеры при принадлежности к премиальному бренду. Методы OLS, CME и DR дают сопоставимые оценки в пределах 36 000 рублей, что указывает на умеренное влияние бренда при прочих равных технических характеристиках. Метод IPW даёт почти вдвое большую оценку, что может быть связано с неоднородностью распределения признаков в группах. Метод DML, напротив, даёт нереалистично высокую оценку, что требует отдельного анализа и, вероятно, технического пересмотра модели.
Вывод: OLS, CME и DR дают устойчивые и согласованные оценки. IPW более чувствителен к дисбалансу и вероятностной модели. Метод DML при неправильной реализации может давать экстремальные результаты. Следует использовать несколько подходов для проверки устойчивости выводов.

Задание 5.5. Локальный эффект воздействия с помощью двойного машинного обучения

Результаты оценок

- **DML без инструмента (на самом деле LATE):**

$$LATE_{DML} = -8\,922.69$$

  Данная оценка была получена как среднее значение индивидуальных эффектов (`effect(X)`) с использованием `LinearDML` (модель воздействия: логистическая

регрессия, модель результата: случайный лес).

- **DML с инструментальной переменной (DMLIV):**

$$LATE_{DMLIV} = -10\,108.99$$

Инструмент: `used_in_cinema`. Метод — `DMLIV` с двумя логистическими моделями для первого шага и случайным лесом для оценки итогового эффекта.

Сравнение и интерпретация

- Оба метода показывают отрицательный локальный эффект: принадлежность к премиальному бренду ассоциирована с уменьшением цены камеры на ∼9–10 тысяч рублей.
- DML без инструмента предполагает отсутствие скрытых переменных, однако может быть подвержен смещению из-за эндогенности.
- Метод с инструментом (DMLIV) фокусируется на эффектах в подгруппе, чувствительной к изменению инструмента (комплаеры), и может быть более причинно интерпретируемым.

Причины различий:

- В отличие от среднего эффекта воздействия (ATE), который усреднён по всей популяции, оценки DML и DMLIV отражают локальные эффекты в определённых подгруппах.
- DML может захватывать корреляции, обусловленные нефакторизуемыми переменными (например, маркетинговыми стратегиями премиальных брендов).
- DMLIV позволяет частично устранить это смещение, однако его результат ограничен интерпретацией в рамках группы комплаеров.

Экономическая интерпретация

- **LATE (без инструмента)** — усреднённый эффект перехода к премиальному бренду среди всей популяции, с предположением корректной спецификации модели.
- **LATE (с инструментом)** — причинный эффект для подгруппы, в которой выбор бренда определяется переменной `used_in_cinema`. Уменьшение цены может отражать смещение спроса в нишевом сегменте (например, профессиональной видеосъёмки), где премиальность воспринимается иначе.

Вывод: Негативные оценки локального эффекта воздействия по обоим методам могут указывать на рыночную специфику премиального бренда: возможно, такие камеры ориентированы на профессиональный сегмент, где цена определяется не только брендом, но и техническими или функциональными требованиями. Метод DMLIV предоставляет более обоснованную причинную интерпретацию, но его результат применим лишь к узкой подгруппе.

Задание 5.6. Оценка условных средних эффектов воздействия (CATE)

Результаты оценки CATE (Conditional Average Treatment Effect)

Оценки среднего эффекта воздействия по каждому методу:

- OLS (по коэффициенту при переменной D): 17 726.45
- S-learner: 27 431.68
- T-learner: 35 964.90
- X-learner: 35 821.81

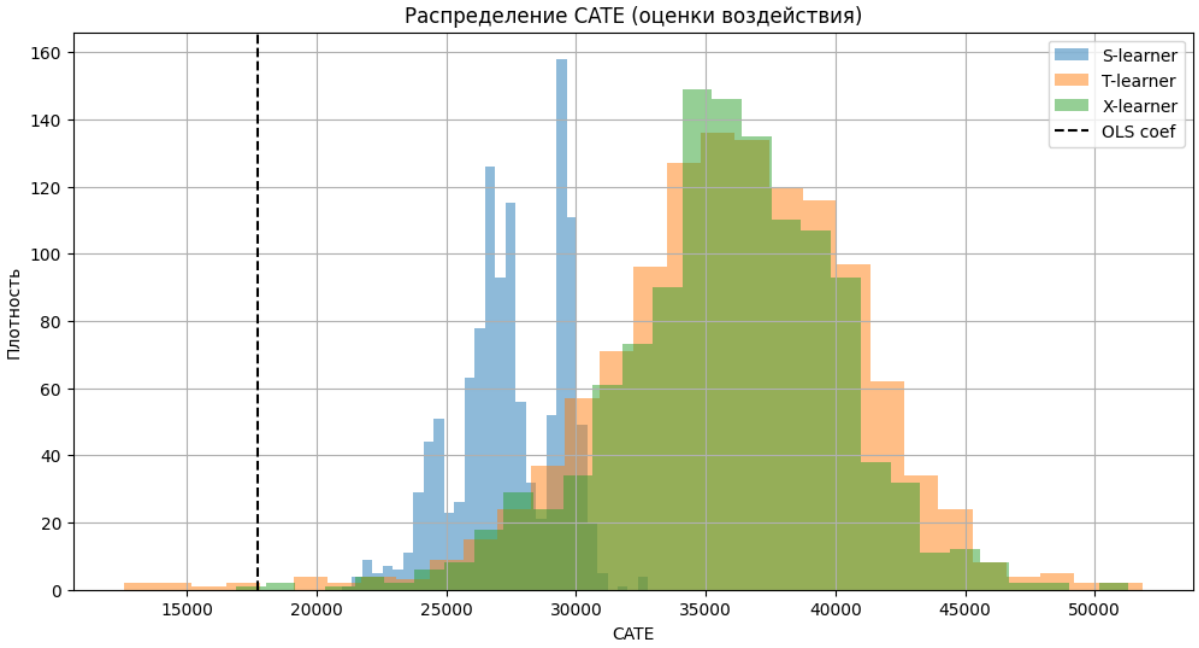

Рис. 6: Распределение оценок CATE по методам: S-learner, T-learner и X-learner. Вертикальная пунктирная линия — оценка OLS.

Сравнение и интерпретация методов

- **OLS** предоставляет глобальную оценку эффекта воздействия, но не учитывает гетерогенность между наблюдениями.
- **S-learner** использует одну модель с включением переменной воздействия D как предиктора. Это простой и интерпретируемый подход, но он может сглаживать гетерогенные эффекты между группами, особенно при сильных различиях.
- **T-learner** строит две модели — отдельно для групп $D = 1$ и $D = 0$. Лучше улавливает гетерогенность, но может страдать от повышенной дисперсии, особенно при несбалансированности классов.
- **X-learner** учитывает асимметрию и использует вероятности принадлежности к группам, чтобы скорректировать предсказания T-learner. Он особенно эффективен при дисбалансе, обеспечивая адаптивные и устойчивые оценки.

Мотивация для X-learner: переменная воздействия (например, премиальность бренда) в нашей выборке несбалансирована (менее 50% единиц). X-learner адаптируется к этой особенности, обеспечивая более устойчивые оценки. Его результат (около 35 800 рублей) близок к T-learner, но обладает преимуществами в плане интерпретации и устойчивости.

Применение в бизнесе и госуправлении

- **Бизнес:** CATE позволяет настраивать маркетинговые стратегии под конкретные сегменты. Например, премиальный бренд может усиливать продвижение в группах с наибольшим индивидуальным эффектом.
- **Госпрограммы:** такие оценки позволяют оценить, где вмешательства (например, сертификация, локализация сборки) наиболее эффективны.
- **Ценообразование:** производители могут использовать CATE для персонализированной ценовой политики.

Дополнительный метод: R-learner (повышенная сложность)

R-learner минимизирует функционал, в котором CATE трактуется как параметр, компенсирующий остатки двух моделей — результата и воздействия. Это обеспечивает гибкость и устойчивость даже при сложных зависимостях.

Преимущества:

- Теоретически строгий подход.

- Устойчивость при слабой корреляции между воздействием и результатом.

Недостатки:

- Требует двойной настройки моделей.
- Повышенная сложность реализации.

Задание 5.7. Сравнение моделей оценки условных эффектов воздействия

Выберите лучшую модель оценивания условных средних эффектов воздействия, используя:

- истинные значения условных средних эффектов (true CATE),
- прогнозную точность на целевой переменной (price),
- псевдоисходы (pseudo outcomes).

Сравнение по метрике PEHE (True CATE)

Модель	**PEHE (True CATE)**
S-learner	4 745.54
T-learner	13 742.21
X-learner	13 447.13

Таблица 4: Сравнение моделей по ошибке PEHE с истинным эффектом

Сравнение по MSE (предсказание цены)

Модель	**MSE (Price pred)**
S-learner	6.95×10^7
T-learner	1.18×10^8
X-learner	1.27×10^8

Таблица 5: Сравнение моделей по ошибке прогноза цены

Сравнение по псевдоисходам (PEHE)

Модель	PEHE (Pseudo Outcome)
S-learner	7 859.23
T-learner	10 508.35
X-learner	11 081.39

Таблица 6: Сравнение моделей по ошибке PEHE на псевдоисходах

Вывод: модель **S-learner** демонстрирует наилучшее качество по всем трём метрикам: минимальная ошибка при оценке истинных эффектов (PEHE), наименьшее среднеквадратичное отклонение при прогнозе цены (MSE), а также наилучшая производительность на псевдоисходах. Это свидетельствует о её устойчивости и точности в контексте данной задачи. Модель **T-learner** проигрывает в точности по всем критериям, тогда как **X-learner** находится между ними, показывая сбалансированные результаты. На практике S-learner может быть предпочтительным методом для индивидуализированной оценки воздействия в условиях ограниченного дисбаланса между группами.