Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Отчет
по дисциплине “Машинное обучение в экономике”
Тема “Оценка влияния наличия сцен 18+ в кинофильмах на кассовые сборы”

Подготовили:
Амзина Полина БЭК221 (семинарист: Погорелова П.В.)
Золотова Елена БЭК222 (семинарист: Зинченко Д.И.)
Стрекаловских Дмитрий БЭК223 (семинарист: Гергенретер А.О.)

Москва, 2025

1. Обоснование темы:

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

Непрерывной или целевой переменной является кассовые сборы кинофильмов в миллионах долларов. Бинарной переменной воздействия является наличие в киноленте сцен, которые предназначаются для взрослой аудитории (насилие, сцены интимного характера, употребление запрещенных веществ и др). Иными словами, присвоение фильму возрастного рейтинга R. Переменная равняется 1, если такой рейтинг был присвоен и 0 в ином случае. Так, целью настоящего исследования является анализ влияния возрастного рейтинга R на кассовые сборы фильмов.

Примечательно, что в исследовании используется возрастной рейтинг, принятый американской ассоциацией MPPA и из-за различий в законодательствах рейтинг R указывает на то, что лицам не достигшим 17 лет требуется обязательное присутствие взрослого, что в России означало бы наличие у фильма маркировки “18+”. По этой причине будем считать, что один год возрастной разницы не является значимым ограничением исследования. Также Министерство культуры Российской Федерации автоматически присваивает иностранным фильма с рейтингом R маркировку “18+”, поэтому будем считать, что они равнозначны.

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.

Так как кассовые сборы являются одним из основных показателей успеха фильма, а для инвесторов ключевым, то изучение влияния наличия определенных сцен в кино может помочь кинопродюсерам и режиссерам, при создании новых кинолент. Так, если зрители негативно реагируют на определенные сцены, то для достижения высокой окупаемости вложений, возможно, стоит избежать их наличия или показывать события в ином виде. Например, при негативной реакции зрителей на употребление ненормативной лексики в дальнейшем возможно использование других лексических оборотов для достижения определенной цели повествования.

Для государств само наличие возрастных ограничений может быть сигналом к ограничению просмотра киноленты в стране (полная блокировка) или частичной

цензуре, что напрямую влияет на качество восприятия киноленты, что является дополнительным сигналом для производственных студий при планировании сценария.

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.

Наличие у фильма возрастного рейтинга R негативно сказывается на кассовых сборах. Во-первых, ограничение аудитории, которая может посетить кинотеатр, приводит к тому, что семьи отказываются от просмотра данных фильмов в пользу тех, которые можно посмотреть с детьми (возрастные рейтинги G, PG и PG-13). Примечательно, что возрастной рейтинг R напрямую не ограничивает аудиторию, а лишь обязывает присутствие взрослого для просмотра киноленты лицами не достигшими 17 лет, однако для семейного просмотра такие фильмы все равно не подходят.

Во-вторых, фильмы с определенными сценами могут иметь сложности с рекламой на определенных площадках (телеканалы, видеохостинги, которые блокируют определенный контент, и др). Также имеет место быть полная блокировка киноленты в некоторых странах, что сильно ограничивает количество потенциальных кинозрителей.

Наше предположение о негативном влиянии подтверждается научной литературой. Так, в работе “Do MPAA Ratings Affect Box Office Revenues?[1]” автор приходит к выводу, что фильмы с рейтингом PG-13 при прочих равных зарабатывают от 13 до 34 миллионов долларов больше, чем фильмы с возрастным рейтингом R. Это объясняется ограничением потенциальной аудитории.

В работе “Does Hollywood Make Too Many R-Rated Movies? Risk, Stochastic Dominance, and the Illusion of Expectation[2]” авторы делают вывод, что Голливуд производит слишком много фильмов с возрастным рейтингом R из-за их экономической неэффективности. Так, R-фильмы менее вероятно получат экстремальные прибыли, но при этом имеют более тяжелые хвосты убытков.

Автор исследования “Film Review Aggregators and Their Effect on Sustained Box Office Performance” приходит к похожему выводу: в долгосрочной перспективе фильмы с возрастным рейтингом R показывают значительно более низкие доходы, чем с

[1] Brooke Conaway, Daniel Ellis Do MPAA Ratings Affect Box Office Revenues? // 2015. - C. 64-88.

[2] Arthur De Vany, W. D. Walls Does Hollywood Make Too Many R-Rated Movies? Risk, Stochastic Dominance, and the Illusion of Expectation // The Journal of Business 75(3). - 2000

рейтингом G/PG. Однако, в краткосрочной перспективе (первые 2 недели после премьеры) возрастной рейтинг не оказывал сильного влияния на кассовые сборы.

1.4 Кратко опишите результаты предшествующих 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.

В работе “Do MPAA Ratings Affect Box Office Revenues?” был проведен анализ 1635 фильмов, выпущенных с 2001 по 2012 года. Для оценки влияния различных факторов с упором на возрастной рейтинг на кассовые сборы, скорректированные на инфляцию, был использован обычный МНК. Работа помимо возрастного рейтинга учитывает такие факторы, как жанр, качество фильма, дистрибьюторская студия, литературные адаптации, наличие сиквелов, время и масштаб релиза, а также профессиональные оценки и использование 3D-технологий. Примечательно, что в основном работа посвящена сравнению фильмов, которые имеют возрастные рейтинги PG, PG-13 и R, то есть внимание не уделяется картинам с рейтингами G и NC-17. Также авторами рассматривается изменение системы рейтингов MPAA и ее влияние на зрительское восприятие.

Для учета эндогенности стратегии релиза авторы оценили 2 модели: “Passion Project Model” (оценка нижних границ эффектов) и “Profit Project Model” (оценка верхних границ эффектов). Первая модель предполагает, что киностудии не всегда идеально продумывают стратегии релиза кинокартин и не учитывают некоторые факторы. Например, в этой модели контролируются переменная, отвечающая за сезонность релиза. Во второй модели предполагается, что все дополнительные факторы, которые могли повлиять на кассовые сборы, были учтены киностудиями при планировании премьеры.

Как было упомянуто ранее, авторы пришли к выводу, что фильмы с рейтингом R зарабатывают на 13–34 миллиона меньше, чем фильмы с более мягкими возрастными ограничениями. Также авторы подчеркивают, что если раньше возрастные рейтинги не были значимыми при прогнозировании кассовых сборов, то спустя время возрастные ограничения стали важнее из-за изменения зрительского поведения.

Авторами настоящей работы отмечено, что к положительным сторонам исследования относятся учет возможной эндогенности, введение новых характеристик фильмов, которые не были исследованы ранее (3D эффекты, адаптации комиксов), а

также контроль смещающих факторов. К недостаткам можно отнести отсутствие данных о бюджете, так как он является ключевым фактором во многих исследованиях, игнорирование конкуренции при оценке стратегий (не учтен одновременный релиз фильмов), нет проверки спецификации модели (возможно зависимость для фильмов нелинейна).

Следующая работа применяет Парето-распределение для доходов и прибыли кинокартин в связи с тем, что данные крайне асимметричны, а также моменты распределения могут быть бесконечными. Принцип первого порядка стохастического доминирования был применен для доходов, бюджета и прибыли кинофильмов. Так, у R фильмов более тяжелый хвост убытков и меньший хвост прибылей в сравнении с фильмами без возрастных ограничений. К положительным сторонам исследование относится классификация фильмов по бюджету и учет макроэкономических колебаний посредством контроля переменных. К неучтенным факторам можно отнести упрощенный подсчет прибыли (50% кассовых сборов – бюджет), а также игнорирование жанров, так один и тот же звездный актер, снимаясь в разных жанрах не может гарантировать коммерческой стабильности. Так, Уилл Смит после премьеры “Люди в черном[3]” в жанре sci-fi и имея колоссальную известность среди зрителей не смог принести того же успеха фильму “После нашей эры[4]” в жанре фэнтези. Также важно учесть, что в выборке присутствует всего 60 фильмов с рейтингом G и 1057 с рейтингом R в связи с чем возможно искажение выводов при сравнении.

Для третьего исследования была использована МНК оценка линейной в логарифмах модели, при этом гетероскедастичность устранялась с помощью робастных ошибок в форме Уайта. Однако данные исследования были сильно ограничены снизу, так выборка состояла из фильмов, входящих в топ 150 США, также для некоторых картин бюджет был оценен приблизительно, что могло повлиять на результаты.

К недостаткам, применяемым нами методами, можно отнести ограниченный набор факторов, которые влияют на кассовые сборы. В большинстве исследований используется около 7–10 различных переменных, в то время как мы ограничиваемся лишь 4 (переменная воздействия и 3 контрольные). К плюсам нашей модели можно отнести:

---

[3] Люди в черном // Кинопоиск URL: https://www.kinopoisk.ru/film/1091/?ysclid=mazpmo0zjm99928730&utm_referrer=yandex.ru (дата обращения: 22.05.2025).

[4] После нашей эры // Кинопоиск URL: https://www.kinopoisk.ru/film/577285/?ysclid=mazpor0ovt398823342 (дата обращения: 22.05.2025).

1) Гетерогенность эффектов: функция кассовых сборов состоит из двух частей (для фильмов с и без рейтинга R), с разными коэффициентами и формой внутри, что достаточно полно описывает влияние разных факторов (бюджета, рейтинга, жанра) на сборы. Это позволит более точно оценить условные средние эффекты CATE и тд.

2) Переменные генерируются через логистическую функцию с взаимодействиями и полиномиальными эффектами, что создает нелинейную зависимость, близкую к реальному миру. Также используются логарифмы, квадраты переменных, перекрестные члены как ratings * budget или $budget^2$.

3) Заданы разные распределения ошибок для фильмов с и без рейтинга R – t-распределение и экспоненциальное, что учитывает дифференциацию разброса сборов для разных категорий фильмов и целевой аудитории, так как кассовые сборы обычно непредсказуемы.

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 основной переменной, которая влияет на кассовые сборы кинофильмов, берется производственный бюджет в миллионах долларов. Влияние бюджета подтверждается практически во всех основных работах, исследующих влияние факторов на кассовые сборы[5]. Переменная является непрерывной.

Также непрерывной переменной, влияющей на кассовые сборы, является зрительская оценка. Зрители при выборе киноленты могут ориентироваться на уже посетивших показ. Так, чем выше оценка кинозрителей, тем выше кассовые сборы.

Бинарной переменной, влияющей на кассовые сборы, является то, является ли фильм сиквелом или он одиночный. Продолжение получают наиболее успешные фильмы, что предопределяет их потенциальный экономический успех в будущем из-за всеобщей любви к героям, желанию узнать продолжение истории или иным причинам. Положительное влияние предыдущих частей на кассовые сборы подтверждается в работе “Success in the Film Industry: What Elements Really Matter in Determining Box-Office Receipts[6]”. Примечательно, что согласно теории упадка[7] фильмы-сиквелы имеют худшие показатели по сравнению с их приквелами. Однако, для настоящего

[5] Topf P. Examining success in the motion picture industry //The Park Place Economist. – 2010. – Т. 18. – No. 15. – С. 1.
[6] Greenaway, M. & Zetterberg B., 2012. Sucess in the Film Industry: What Elements Really Matter in Determining Box-Office Receipts
[7] Edward P. Lazear, The Peter Principle: A Theory of Decline

исследования предполагается, что выборка генерируется за определенный год, что означает, что фильмы сиквелы сравниваются с только вышедшими одиночными фильмами, а не со своими приквелами.

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

Инструментальной переменной является переменная на возрастной рейтинг предыдущих фильмов, то есть переменная равняется 1, если предыдущие работы режиссера получали возрастной рейтинг R, и 0 иначе. Переменная выбрана из-за особенностей присвоения возрастного рейтинга в США. Так, после создания фильма создается специальная комиссия, которая определяет возрастной рейтинг. В случае неудовлетворительного ответа возможна либо подача апелляции, либо удаление определенных сцен, что для режиссера крайне нежелательно. Наличие в фильмографии фильмов, ориентированных на взрослую аудиторию, может означать принципиальность режиссера в сохранении всех сцен фильма. Например, все фильмы Квентина Тарантино имеют возрастной рейтинг R, и если он решит выпустить еще один фильм, то наиболее вероятно, что новая картина тоже будет иметь возрастной рейтинг R. Напротив, Альфонсо Куарон в своей фильмографии имеет картины абсолютно всех возрастных рейтингов, поэтому нельзя предсказать то, с каким возрастным рейтингов будет его следующая работа.

При этом, предыдущие возрастные рейтинги кинокартин режиссера не влияют напрямую на кассовые сборы текущего фильма, так как зрители выбирают фильм по текущим характеристикам, а не наличию определенных сцен в прошлом.

Так, настоящий инструмент является релевантным и валидным.

1.7 В случае необходимости приведите дополнительные содержательные комментарии о целях, задачах, методологии и вкладе вашего исследования.

Заметим, что с одной стороны фильмы-сиквелы подразумевают наличие определенных зрителей из-за любви к предыдущей части, а с другой, согласно теории упадка фильмы-сиквелы хуже фильмов-приквелов. Так, даже если сиквелы являются наиболее кассовыми фильмами в определенный период, им все равно не удается побить рекорд первой части. Согласно этому, введем предположение о том, что в работе генерируются характеристики для фильмов одного года выпуска. Так, фильмы сиквелы будут соревноваться с приквелами других франшиз, что нивелирует результат теории упадка.

Также предположим, что бюджет является ненаблюдаемой переменной, так как зачастую кинокомпании его скрывают по личным соображениям. Это предположение будет использовано в дальнейшем.

2. Генерация и предварительная обработка данных:

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

В работе используются следующие обозначения: Y — кассовые сборы фильма (зависимая переменная, непрерывная), D — наличие рейтинга R (переменная воздействия, бинарная, равна 1, если присутствует факт наличия сцен 18+, 0 иначе), Z — инструментальная переменная: опыт режиссёра с R-жанром (1, если ранее снимался в жанрах, содержащих R рейтинг, 0 иначе), X = (sequel, ratings, budget) — вектор контрольных переменных.

В данном пункте происходит построение модели и генерация данных для анализа влияния возрастного рейтинга R на кассовые сборы фильма. В отличие от тривиальных моделей, мы учли вероятностную природу формирования как инструментальной, так и эндогенной переменной, ввели нелинейности, взаимодействия переменных и гетероскедастичные ошибки.

На первом этапе смоделирован инструмент, отражающую жанровую историю режиссера — снимал ли он ранее фильмы с рейтингом R. Эта переменная не задавалась случайно, а определялась через логистическую функцию от нескольких характеристик фильма: рейтинга пользователей, того, является ли фильм сиквелом, бюджета, квадратичного эффекта бюджета, а также взаимодействия между рейтингом и бюджетом. Это позволило отразить реальное поведение режиссёров, которые склонны снимать фильмы с рейтингом R в зависимости от жанровой специфики и коммерческого контекста.

Далее моделировалась эндогенная переменная — факт наличия рейтинга R. Здесь используется логистическая функция вероятности, но с учетом инструментальной переменной и нелинейной зависимости от бюджета, оценки пользователей, логарифма рейтингов, квадрата бюджета и взаимодействия между рейтингом и инструментальной переменной. Рейтинг R не случайный — он зависит от художественного и производственного профиля фильма (эта зависимость явна видна в формуле), тем самым показываем структуру, в которой инструментальная переменная влияет на рейтинг R, но не напрямую на сборы, что даёт нам возможность провести оценку с поправкой на эндогенность.

На третьем этапе мы моделировали кассовые сборы, задав две разные функции дохода: одну — для фильмов с рейтингом R, другую — без. Эти функции зависят от рейтинга пользователей, бюджета, сиквела и инструментальной переменной, но при этом между двумя группами фильмов заложены различные коэффициенты и функциональные формы, так как имеют различные эффекты воздействия на результат. Например, логарифмические и квадратичные трансформации рейтингов и бюджета отражают убывающую или возрастающую отдачу от этих факторов. Влияние инструментальной переменной входит в обе функции дохода, моделируя возможное жанровое влияние на коммерческий результат. Кроме того, креативной частью модели является введение разных распределений ошибок для двух групп: t-распределения с тяжелыми хвостами для фильмов без рейтинга R и смещенного экспоненциального распределения для фильмов с рейтингом R. Это позволяет учесть несимметричную, гетероскедастичную природу случайных факторов, влияющих на сборы, что делает модель ближе к реальности.

Этап 1: Генерация инструментальной переменной Z. Это вероятностная модель:

$$P(Z = 1|X) = \log i\,stic\left(\frac{1}{2} * m\right), \text{ где}$$

$$m = 0.3 * ratings + 1.5 * sequel - 4 + e_1$$

$$P(Z = 1|X) = \frac{1}{(1 + e^m)}$$

m – инструмент, жанры, где режиссер снимался раньше. e_1 является шумом с нормальным стандартным распределением для реалистичности вероятности. Режиссёры, ранее снимавшие сиквелы, популярные фильмы или с большим бюджетом, чаще пробовали себя в жанрах с рейтингом R.

Этап 2: Эндогенность D — формирование рейтинга R. Вероятность получения рейтинга R зависит от жанровой истории режиссёра (инструмент), бюджета, оценок, и их взаимодействий:

$$P(D = 1|X) = \log i\,stic(2.2 * Z + 0.01 * budget + 0.1 * ratings + 0.001 * budget^2 + 0.07 * \log(ratings + 1) + 0.002 * Z * ratings - 8)$$

$$q = 2.2 * Z + 0.01 * budget + 0.1 * ratings + 0.001 * budget^2 + 0.07 * \log(ratings + 1) + 0.002 * Z * ratings - 8$$

$$P(D = 1|X) = \frac{1}{(1 + e^q)}$$

Наличие рейтинга R зависит от контента и стиля режиссёра, а также от качества фильма и вложенных средств.

Этап 3: Ошибки модели: ошибки различаются, то есть модель гетероскедастична и несимметрична, что делает её реалистичной и нелинейной.

$\varepsilon_0 \sim t_{10} * 8$ — тяжелохвостое распределение ошибок для фильмов без рейтинга R

$\varepsilon_1 \sim Exp\,(25) - 10 + 0.001 \cdot budget$— асимметричное распределение ошибок для фильмов с рейтингом R

Этап 4: Функции дохода.

Для фильмов с рейтингом R (D=1):

$$g_1 = -2 * \log(ratings + 1) + 1.0 * budget + 12 * sequel - 0.03 * budget^2 + 0.3 * Z$$

Для фильмов без рейтинга R (D=0):

$$g_0 = -1.5 * \log(ratings + 1) + 1.3 * budget + 10 * sequel + 0.017 * ratings2 + 0.6 * Zg0$$
$$= -1.5 * \log r\,atings + 1 + 1.3 * budget + 10 * sequel + 0.017 * ratings2 + 0.6 * Z$$

Общая модель кассовых сборов:

$$Y = D * (g_1 + e_1) + (1 - D) * (g_0 + e_0) + 140$$

Добавление константы 140 — задаёт минимальный ожидаемый уровень сборов (средняя дистрибуция в кинотеатрах).

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

1) Переменная воздействия D (рейтинг R): влияние на сборы нелинейно и неоднозначно.

Позитивное влияние через разные ниши (niche targeting): фильмы с рейтингом R могут иметь высокий спрос у определенных потребителей контента (например, ужастики, экшн). Негативное эффект через ограничение аудитории: меньшее количество зрителей (до 17 лет) может просматривать такие фильмы, поэтому отрицательный коэффициент. О чем говорит разная функциональная форма g1 и g0: разные коэффициенты, разное влияние бюджетов и оценок.

2) Инструментальная переменная Z (жанровый опыт режиссёра)

Ожидание положительного влияния на рейтинг R, но не напрямую на сборы, если инструмент валиден (далее будет проверено). Через g0 и g1 может иметь влияние на доходы — отражает косвенную жанровую специфику. Коэффициенты +0.3*Z в g1 и +0.6*Z в g0 (лучше монетизируется высокий опыт режиссера) показывают возможную

слабую корреляцию между жанровым опытом и вкусами аудитории. Фильмы «от режиссера хорроров» могут иметь специфичную аудиторию даже без высоких рейтингов. Полная экзогенность нарушается, и в реальных данных идеально валидных инструментов почти не существует. Это создает более честный тест для методов оценки. Задача была обеспечить реалистичную гетерогенность эффекта: один и тот же опыт может по-разному влиять на кассу в зависимости от того, есть ли у фильма рейтинг R.

3) Контрольные переменные X

Sequel: положительное влияние на сборы: присутствует часть людей, которые ранее были фанатами и готовы прийти посмотреть продолжение фильма, маркетинг осуществляется быстрее и имеет более сильный эффект. Коэффициенты +12 и +10 в g1 и g0.

Ratings (оценки пользователей). Влияние нелинейное: -log(ratings+1) снижает эффект плохих оценок, ratings² усиливает хорошее восприятие. Критически плохие оценки уменьшают сборы, так как снижают лояльность новой аудитории, но более высокие помогают увеличить шансы на высокие охваты и массовость просмотра.

Budget: большие бюджеты обычно ассоциированы с высокими сборами, эта связь подтверждается во многих исследованиях. Чрезмерный бюджет может давать уменьшение отдачи — и в модели есть криволинейность (через квадрат и взаимодействия).

2.3. 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 или больше 0.9 ни для одной из бинарных переменных.

Ниже приведены три таблицы (1 – корреляционная матрица, 2 – описательные статистики для непрерывных переменных, 3 – для бинарных). А также для исследовательского интереса было визуализировано распределение кассовых сборов. Условия соблюдаются: доля единиц выполняется для всех бинарных переменных - например, доля для sequel = 0.3 (см.Табл.3), а число наблюдений больше 1000.

Самая высокая корреляция наблюдается у revenue и budget (см.Табл.1) равная 0.34, что говорит о логичной интерпретируемости, где кассовые сборы и бюджет положительно зависят друг от друга. Умеренная корреляция у кассовых сборов с

инструментальной переменной, показывающая более высокие сборы фильмов, снятые режиссерами в аналогичном жанре. Сиквел положительно коррелирует с revenue, что подтверждает гипотезу об успешности продолжения известных франшиз. Остальные переменные имеют более низкую связь с кассовыми сборами и между собой.

Ориентируясь на средние значения выручки в 244 млн (см.Табл.2), заметно, что разброс умеренный и колеблется от 187 до 452 млн (см.Рис.1) и подтверждает нормальность структуры данных. В свою очередь, значение рейтинга равное 6.5 указывает на соответствие симулированных данных изученным статьям и реальной жизни для массового кинопроката.

	revenue	rating_r	sequel	ratings	budget	director_past_genre
revenue	1.00	0.02	0.21	-0.02	0.34	-0.02
rating_r	0.02	1.00	0.04	0.04	0.46	0.24
sequel	0.21	0.04	1.00	0.00	0.00	0.16
ratings	-0.02	0.04	0.00	1.00	-0.02	0.06
budget	0.34	0.46	0.00	-0.02	1.00	-0.01
director_past_genre	-0.02	0.24	0.16	0.06	-0.01	1.00

Табл. 1. Корреляционная матрица между переменными.

	revenue	ratings	budget
mean	244.322412	6.518583	87.839707
std	24.927627	1.007164	9.953730
50%	240.148695	6.517282	87.785386
min	187.819689	2.980575	51.251973
max	451.915922	10.000000	128.646299

Табл. 2. Описательные статистики для непрерывных переменных.

	variable	доля	количество
0	rating_r	0.786	7856
1	director_past_genre	0.330	3299
2	sequel	0.300	3005

Табл. 3. Описательные статистики для бинарных переменных.

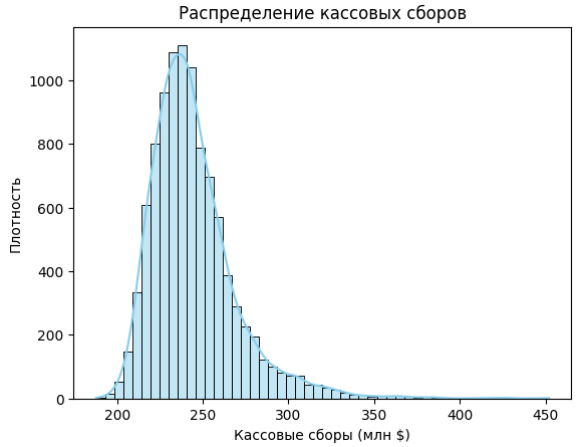


Рис. 1. Распределение кассовых сборов

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

Нами была поделена выбора на тестовую, в нее было решено включить 25% наблюдений и обучающую 75% от всех.

Размер обучающей выборки: (7500, 6)

Размер тестовой выборки: (2500, 6)

2.5. В случае необходимости проведите дополнительный анализ и приведете дополнительные комментарии о процессе генерации данных, описательных статистиках и т.д.

Для проверки мультиколлинеарности помимо построения корреляционной матрицы (где подозрительной корреляции выше 0.7 между разными переменными обнаружено не было) из прошлых пунктов была построена таблица со значениями VIF (см. Табл 4). Так как результаты оказались меньше 10, то проблемы мультиколлинеарности обнаружено не было, следовательно дополнительного анализа для устранения проводить не потребовалось.

	переменная	VIF
0	const	133.040437
1	budget	1.461462
2	rating_r	1.295979
3	revenue	1.156855

Табл. 4. Значения VIF для переменных

Был проведен дополнительный анализ для проверки инструмента на релевантность и валидность.

Релевантность через F-статистику первого шага 2МНК (IV). F-test: F=816.59, p=0.00. Результат: F-статистика больше 10, p-value околонулевое, что говорит о релевантности. Инструмент сильный. Валидность: IV-регрессия: зависимая переменная — кассовые сборы, эндогенная — R рейтинг, инструмент — Z Было построено две регрессии (IV и OLS), которые в последующем сравнивались (см. Табл 5).

Model Comparison

	OLS	IV
Dep. Variable	revenue	revenue
Estimator	OLS	IV-2SLS
No. Observations	10000	10000
Cov. Est.	robust	robust
R-squared	0.1835	0.1824
Adj. R-squared	0.1832	0.1821
F-statistic	2273.5	2198.9
P-value (F-stat)	0.0000	0.0000
Intercept	158.33	156.08
	(60.632)	(45.404)
budget	1.0440	1.0865
	(39.869)	(21.207)
ratings	-0.1365	-0.0949
	(-0.6033)	(-0.4187)
sequel	11.910	11.984
	(24.138)	(23.887)
rating_r	-10.701	-12.950
	(-23.691)	(-5.5092)
Instruments		director_past_genre

Табл. 5. Сравнительная таблица двух результатов двух регрессий IV и OLS.

Дополнительные тесты показали, что наш инструмент — валидный (не коррелирует с ошибками) и релевантный (значимо влияет на rating_r).

3. Классификация

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.

На данном этапе производится отбор признаков для прогнозирования переменной воздействия (D). Цель — построить модель, оценивающую условную вероятность получения воздействия при заданных характеристиках P(D=1|X). В качестве признаков X следует выбирать переменные, которые могут одновременно влиять как на переменную воздействия (рейтинг R), так и на целевую переменную (кассовые сборы revenue).

Для прогнозирования наличия рейтинга R (rating_r) были отобраны следующие признаки:

- **sequel (контрольная переменная):** Продолжения фильмов (сиквелы) в большинстве случаев наследуют рейтинг и целевую аудиторию от оригинальных частей серии, делая этот признак важным предиктором.
- **ratings (контрольная переменная):** Средние пользовательские оценки могут косвенно указывать на контент фильма (например, более взрослый или нишевый контент может получать специфические оценки) и, следовательно, коррелировать с возрастным рейтингом.
- **budget (контрольная переменная):** Бюджет фильма часто определяет его масштаб, жанр и маркетинговую стратегию. Высокобюджетные блокбастеры чаще ориентированы на широкую аудиторию (и более мягкий рейтинг), тогда как фильмы с меньшим бюджетом могут быть более рискованными и ориентированными на взрослую аудиторию.

3.2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов:

• на обучающей выборке.

• на тестовой выборке.

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

Проинтерпретируйте полученные результаты.

Для решения задачи были выбраны три модели с различными принципами работы:

- **Логистическая регрессия:** Классический линейный метод, который моделирует логарифм шансов $\log\left(\frac{p}{1-p}\right)$ бинарной целевой переменной как линейную комбинацию признаков. Модель легко интерпретируема, быстра в обучении и эффективна при наличии линейной зависимости. Гиперпараметры: c - гиперпараметр регуляризации
- **Случайный лес (Random Forest):** Ансамблевый метод, основанный на построении множества решающих деревьев на различных подвыборках данных. Итоговый прогноз усредняется, что значительно снижает риск переобучения и повышает устойчивость. Метод способен улавливать сложные нелинейные зависимости. Гиперпараметры: n_estimators - число деревьев, max_depth - максимальная глубина дерева
- **Метод k-ближайших соседей (KNN):** Простой непараметрический метод, который классифицирует объект на основе класса большинства его k соседей в пространстве признаков. Метод не делает предположений о распределении данных, но его производительность чувствительна к масштабу признаков, шуму и выбору метрики расстояния. Гиперпараметры: n_neighbors - число ближайших соседей

Оценка качества моделей проводилась по метрике точности (Accuracy) на обучающей, тестовой выборках и с помощью кросс-валидации. Сравнение этих показателей позволяет оценить устойчивость и склонность моделей к переобучению: если точность на обучении сильно больше точности на тесте, то модель переобучена; если все три показателя близки, модель считается устойчивой.

Модели были обучены со случайно выбранными гиперпараметрами (LogReg: c=1; Random Forest: n_estimators=100, max_depth=5; KNN: n_neighbors=5).

Model	**Train Accuracy**	**Test Accuracy**	**CV Accuracy**
Logistic Regression	0.820	0.811	0.819
Random Forest	0.838	0.808	0.819
KNN	0.855	0.781	0.799

Таблица 6. Сравнение точности моделей с исходными гиперпараметрами

Согласно результатам (см. Табл. 6) все три модели показывают схожую и достаточно высокую точность на кросс-валидации и тестовой выборке. У KNN наблюдается заметный разрыв между точностью на обучающей (85.5%) и тестовой (78.1%) выборках, что указывает на небольшое переобучение. Логистическая регрессия и случайный лес демонстрируют более стабильные результаты, что говорит о их лучшей обобщающей способности на данном этапе.

3.3 Для каждого метода с помощью кросс-валидации на обучающей выборке под берите 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.

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

Для улучшения качества моделей был выполнен подбор оптимальных гиперпараметров с помощью **GridSearchCV** на обучающей выборке. Этот метод осуществляет исчерпывающий перебор всех возможных комбинаций заданных гиперпараметров и с помощью кросс-валидации находит ту, которая обеспечивает наилучшее значение целевой метрики (в данном случае — Accuracy).

Для случайного леса дополнительно был применен подбор по **Out-of-Bag (OOB) ошибке**. OOB-ошибка — это внутренняя оценка обобщающей способности модели, рассчитываемая на объектах, не попавших в бутстрэп-выборку для каждого дерева. Это позволяет оценивать модель без необходимости создавать отдельную валидационную выборку или использовать кросс-валидацию, что экономит вычислительные ресурсы.

Как можно заметить тюнинг незначительно улучшил показатели моделей, что говорит об их изначальной стабильности. KNN получил небольшой прирост точности после подбора гиперпараметров (см. Табл. 7).

Модель	**Гиперпараметры (до)**	**CV acc (до)**	**Test acc (до)**	**Гиперпараметры (после)**	**CV acc (после)**	**Test acc (после)**
Logistic Regression	C=1	0.819	0.811	C=0.01	0.820	0.810
Random Forest	n_estimators=100; max_depth=5	0.819	0.808	n_estimators=100; max_depth=5	0.819	0.808
KNN	n_neighbors=5	0.799	0.781	n_neighbors=9	0.806	0.791

Таблица 7. Результаты тюнинга гиперпараметров по метрике Accuracy

Оба метода привели к выбору схожих гиперпараметров и практически идентичной точности на тестовой выборке (см. Табл. 8). OOB-оценка является более быстрым аналогом кросс-валидации для случайного леса, но может быть менее стабильной при малом числе деревьев.

Критерий	**n_estimators**	**max_depth**	**CV acc**	**OOB score**	**Test accuracy**
GridSearchCV	100	5	0.819	—	0.808
OOB	200	5	—	0.819	0.806

Таблица 8. Сравнение результатов тюнинга случайного леса по OOB-ошибке и кросс-валидации (GridSearchCV)

3.4. Повторите предыдущий пункт, используя любой альтернативный критерий качества модели. Обоснуйте возможные преимущества и недостатки этого альтернативного критерия. Повышенная сложность: дополнительно самостоятельно запрограммируйте не представленный в стандартных библиотеках критерий качества и используйте его для тюнинга гиперпараметров. Сравните результат стандартного и вашего критериев, а также опишите его

преимущества и недостатки: в каких случаях его разумно применять, а в каких случаях он может оказаться не очень полезен.

Точность (Accuracy) не всегда является лучшей метрикой, особенно при несбалансированных классах. Были рассмотрены два альтернативных критерия:

- **F1-score:** Гармоническое среднее между точностью (precision) и полнотой (recall), $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$. F1-score является хорошим выбором, когда важен баланс между ложноположительными (FP) и ложноотрицательными (FN) ошибками.
- **Собственный критерий:** была предложена взвешенная метрика $CustomScore = 0.7 \cdot recall + 0.3 \cdot precision$. Такой критерий можно использовать, когда минимизация ложноотрицательных ошибок (пропуск фильмов с рейтингом R) является более приоритетной задачей, чем минимизация ложноположительных.

Модель	**Лучшие параметры (F1)**	**CV F1**	**Train F1**	**Test F1**
Logistic Regression	{'C': 0.01}	0.894	0.893	0.888
Random Forest	{'max_depth': 5, 'n_estimators': 50}	0.893	0.898	0.888
KNN	{'n_neighbors': 9}	0.883	0.901	0.873

Таблица 9. Результаты тюнинга гиперпараметров по метрике F1-score

Результаты тюнинга показывают, что все модели показывают высокие и близкие значения F1-score. Логистическая регрессия и случайный лес немного превосходят KNN на тестовой выборке (см. Табл. 9).

Согласно результатам тюнинга по собственному критерию (см. Табл. 10) случайный лес показал наилучшие и наиболее стабильные результаты по данному критерию. Использование собственного критерия позволяет гибко настраивать модель под специфические задачи.

Модель	Лучшие параметры (Custom)	CV Custom	Train Custom	Test Custom
Logistic Regression	{'C': 0.01}	0.925	0.922	0.923
Random Forest	{'max_depth': 3, 'n_estimators': 200}	0.932	0.933	0.930
KNN	{'n_neighbors': 9}	0.902	0.920	0.894

Таблица 10. Результаты тюнинга гиперпараметров по собственному критерию (Custom Score)

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

Повышенная сложность: дополнительно выполните это задание для Байесовской сети и сравните ее ROC-кривую и AUC с теми, что были получены для иных методов.

ROC-кривая (Receiver Operating Characteristic) — это график, показывающий зависимость доли истинно положительных классификаций (TPR) от доли ложно положительных классификаций (FPR) при изменении порога принятия решения. **AUC (Area Under the Curve)** — площадь под ROC-кривой, которая служит интегральной мерой качества модели, независимой от порога. Чем ближе AUC к 1, тем лучше модель различает классы.

Дополнительно была построена **Байесовская сеть** — вероятностная графическая модель, представляющая зависимости между переменными в виде ориентированного ациклического графа (DAG) со структурой:
sequel → rating_r

ratings → rating_r

budget → rating_r.

Логистическая регрессия и случайный лес показывают наилучшее качество (AUC ≈ 0.82), что говорит об их высокой способности к различению классов (см. Табл. 11). Байесовская сеть демонстрирует хороший результат, незначительно уступая

лидерам. KNN показывает самый низкий AUC. Все модели значительно превосходят случайное угадывание (AUC = 0.5).

Модель	**AUC**
Logistic Regression	0.818
Random Forest	0.813
Bayesian Network	0.803
KNN	0.754

Таблица 11. Сравнение моделей по метрике AUC на тестовой выборке

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 пред ставьте в форме таблицы, в которой должны быть указаны как AUC, так и прибыли (на тестовой выборке). Проинтерпретируйте полученный результат.

Повышенная сложность: предложите, содержательно обоснуйте и примени те собственную, отличную от линейной функцию прибыли от прогнозов.

Матрица ошибок показывает детальное распределение верных и неверных классификаций (TP, TN, FP, FN). На ее основе можно рассчитать бизнес-метрики, такие как прибыль. Была предложена линейная функция прибыли, где каждому исходу присвоена "цена":

- TP (Истинно-положительный): +100
- TN (Истинно-отрицательный): +10
- FP (Ложно-положительный): -20
- FN (Ложно-отрицательный): -50

Цель — найти для каждой модели оптимальный порог классификации, который максимизирует эту функцию прибыли на обучающей выборке. Этот подход позволяет адаптировать модель к асимметричной стоимости ошибок.

Согласно результатам поиска оптимального порога и итоговой прибыли на тестовой выборке (см. Табл. 12), логистическая регрессия и случайный лес показали одинаковую максимальную прибыль. Важно отметить, что оптимальные пороги

значительно ниже стандартного значения 0.5. Это связано с тем, что штраф за пропуск фильма с рейтингом R (FN = -50) выше, чем за ложное его обнаружение (FP = -20), поэтому модели выгодно смещать порог для минимизации более "дорогих" ошибок.

Модель	**Оптимальный порог**	**AUC**	**Прибыль (тест)**
Logistic Regression	0.14	0.818	183310
Random Forest	0.34	0.813	183070
KNN	0.12	0.754	182650

Таблица 12. Результаты максимизации прибыли на тестовой выборке

Собственная функция прибыли

Для более реалистичного моделирования была предложена нелинейная функция прибыли, которая учитывает эффект "насыщения" или "растущей цены ошибки":

- **За каждый TP**: +100, но если общее число TP превышает 1800, то за каждый последующий +50.
- **За каждый FP**: -20, но если общее число FP превышает 500, то за каждый последующий -100.
- **TN и FN**: Стоимость остается прежней (+10 и -50 соответственно).

Такой подход позволяет моделировать ситуации, где, например, начальные ложные срабатывания наносят небольшой ущерб, но их большое количество становится критичным.

Применение нелинейной функции прибыли привело к существенному изменению оптимальных порогов — они стали выше, чем при линейной функции (см. Табл. 13). Это указывает на то, что модель, стремясь избежать больших штрафов за многочисленные FP, стала более "осторожной".

Модель	Оптимальный порог	Прибыль (тест)
Logistic Regression	0.59	169010
Random Forest	0.56	169150
KNN	0.45	166270

Таблица 13. Результаты максимизации прибыли с помощью собственной функции на тестовой выборке

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.

Структуру Байесовской сети (DAG) можно задать экспертно, на основе априорных знаний, или обучить на данных с помощью специальных алгоритмов. Был использован алгоритм **HillClimbSearch** со скоринговой функцией BIC, который итеративно изменяет структуру графа (добавляя, удаляя или изменяя направление ребер) в поиске той, которая наилучшим образом описывает данные.

- **Предполагаемый DAG:** Была задана простая структура, где все признаки напрямую и независимо влияют на целевую переменную rating_r.
- **Обученный DAG:** Алгоритм HillClimbSearch выявил более сложную структуру с дополнительными связями между признаками, например, ('rating_r', 'sequel'), который, однако, совсем не поддается логике.

Сравнение точности на тестовой выборке (см. Табл 14):

Модель Байесовской сети	AUC
Предполагаемый DAG (ручная структура)	0.803
Обученный DAG (HillClimbSearch)	0.797

Таблица 14. Сравнение качества прогнозов Байесовской сети с ручной и обученной структурой

Обученная структура показала практически идентичное качество в сравнении с упрощенной ручной структурой. Это говорит о том, что для данной задачи учет дополнительных взаимосвязей между предикторами не привел к значительному улучшению прогнозов, и исходные предположения о прямом влиянии признаков были достаточно точны.

3.8. На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор. Примечание: необходимо самостоятельно сформулировать разумный критерий, в соответствии с которым будут определяться лучшая и худшая модели.

Для выбора лучшей и худшей модели был использован **AUC на тестовой выборке** как основной критерий, поскольку он комплексно оценивает разделительную способность модели и устойчив к дисбалансу классов. Дополнительно учитывался **F1-score**, так как он важен для задач с неравной стоимостью ошибок.

Модель	**AUC**	**Test F1**
Logistic Regression	**0.818**	**0.888**
Random Forest	0.813	**0.888**
Bayesian Network	0.803	-
KNN	0.754	0.873

Таблица 15. Итоговое сравнение классификаторов по ключевым метрикам

- **Лучший классификатор: Логистическая регрессия.** Модель продемонстрировала наибольшее значение по AUC, однако по F1-score на тестовой выборке значение сравнялось со случайным лесом. Несмотря на это, она также показала высокую стабильность результатов, что свидетельствует о ее превосходной и надежной предсказательной способности в данной задаче.
- **Худший классификатор: Метод k-ближайших соседей (KNN).** Модель показала самые низкие результаты по всем ключевым метрикам (AUC, F1-score, Прибыль), а также была склонна к переобучению в первоначальном анализе.

3.9. 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. Добавление обычной регуляризации к классическим методам не считается отдельным методом.

В анализ был включен **CatBoost** — современный алгоритм градиентного бустинга, не входящий в стандартную библиотеку scikit-learn. Его ключевые особенности:

- **Ordered Boosting:** уникальная техника построения деревьев для борьбы с переобучением.
- **Встроенная обработка категориальных признаков:** один из самых эффективных в индустрии механизмов работы с нечисловыми данными.
- **Симметричные деревья:** ускоряют обучение и предсказание.

CatBoost известен своей высокой точностью и эффективностью на табличных данных. CatBoost был обучен и настроен с помощью GridSearchCV.

Модель	**Test AUC**
Logistic Regression (лучшая)	**0.818**
CatBoost (тюнингованный)	0.817

Таблица 16. Сравнение AUC лучшей модели (Логистическая регрессия) и CatBoost

Тюнингованная модель CatBoost показала очень высокий результат, практически сравнимый с логистической регрессией, но все же незначительно уступила ей (см. Табл. 16). Это подтверждает, что для данной конкретной симуляции данных более простая линейная модель оказалась наиболее эффективной. Тем не менее, CatBoost остается мощным инструментом, который мог бы показать лучшие результаты при наличии более сложных нелинейных зависимостей в данных.

3.10. В случае необходимости проведите дополнительный анализ.

Несколько идей как можно улучшить проделанную работу:

1. **Анализ признаков.** Если изучить коэффициенты моделей, то будет легче понять, какие признаки и в какой степени влияют на прогноз. Это добавит работе

интерпретируемости и объяснит, почему модель работает (хоть и на симулированных данных).

2. **Анализ ошибок модели.** Можно проанализировать фильмы, которые модели классифицировали неверно, чтобы найти в них общие закономерности и таким образом сделать более детальный алгоритм для анализа моделей

3. **Проверка стабильности результатов.** Имеет смысл проверить устойчивость выводов, повторив анализ несколько раз с разным сидом генерации и разбиением данных. Это покажет, является ли выбор лучшей модели устойчивой закономерностью, что докажет надежность заключений.

4. Регрессия

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.

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

Отбор признаков для прогнозирования: контрольные переменные (сиквел, бюджет и рейтинг) - x признаки, y - целевая переменная (кассовые сборы). Переменная воздействия не включена, чтобы не было смещения оценок.

Переменная сиквел является категориальной, которая влияет на интерес аудитории и предсказанные значения кассовых сборов в регрессии. Бюджет фильма высоко коррелирует с кассовыми сборами и рейтингом, из-за прямого влияния на качество съемки, монтажа, рекламу. Рейтинг выражен в числовом формате, отражает различные восприятия профессиональных критиков или зрителей с точки зрения успеха в киноиндустрии, что влияет на текущие кассовые сборы или дальнейшие (в ближайшие дни, недели, пойдут ли люди смотреть).

Бинарная переменная помогает точнее смоделировать структурные сдвиги, эффекты и наличие фиктивных переменных. Неучет сиквела может привести к пропуску важной переменнй и смещению оценок, так как он влияет на спрос, другие переменные и затем кассовые сборы. Бюджет напрямую связан с затратами на производство и исключение данной переменной привело бы к неверной интерпретации влияния других факторов на кассовые сборы, эффект был бы сильнее или ниже. Бюджет часто включается в анализ и является экзогенной переменной. Рейтинг является фактором общественного успеха и признания, одобрения социумом, что отражает репутацию производителя и режиссера. В современной литературе вегда учитываются данные опросов, мнения целевой группы для правильной оценки влияния данных.

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 (используйте только обучающую выборку). Проинтерпретируйте полученные результаты.

3 метода: для МНК, случайного леса и градиентного бустинга. RMSE - среднеквадратичная ошибка, чувствительная к выбросам, учитывает крупные ошибки. Считается как корень из $\frac{1}{n \cdot \text{СУММ}\,(y_i - y^{\sim})^2}$, то есть разницу в квадрате между истинным и предсказанным значением. Мы хотим найти RMSE как можно меньше. MAPE - средняя ошибка в процентах, считается как $\frac{1}{n \cdot \text{СУММ}\,\frac{|y_i - y^{\sim}|}{y_i}} \cdot 100\%$, то есть насколько процентов модель в среднем ошибается относительного истинного значения. Чувствительна к нулевым значениям и стремимся ошибку минимизировать, относительный параметр.

Цель: построить регрессионную модель, которая как можно более точнее предсказывает значение целевой переменной (кассовые сборы) на основе объясняющих. Далее будем оценивать качество модели с помощью разных методов.

	RMSE обучающая	RMSE тестовая	RMSE CV	MAPE train	MAPE test	CV MAPE
Линейная модель	22.86	22.87	22.85	6.67%	6.7%	6.67%
Случайный лес	22.33	22.93	22.91	6.56%	6.69%	6.69%
Градиентный бустинг	22.07	22.99	23.05	6.48%	6.71%	6.73%

Табл. 17. Значение разных метрик для трех анализируемых моделей

1) Линейная модель МНК, результаты: значение средней квадратичной ошибки меньше всего на обучающей выборке, однако практически идентично с кросс-валидацией. Средняя ошибка в процентах также ниже для обучающей выборки и совпадает на кросс-валидации, однако значения примерно схожи и модель стабильна, нет переобучения. Данный метод МНК минимизирует сумму квадратов отклонений и наиболее удобен для линейной модели, но в нашем случае это не лучший вариант, так как присутствуют квадраты в функциях, не хватает сложных зависимостей и их интерпретации при МНК. MAPE 6.7% говорит о том, что модель в среднем на 6.7% ошибается от истинного значения.

2) Случайный лес состоит из ансмаблевого метода с множеством деревьев, из-за чего устойчив к выбросам и не реагирует на мультиколлинеарность. В среднем результаты улучшились относительно МНК оценки, в RMSE значение обучающего ниже тестового на 6 десятых, что говорит о небольшом переобучении, но нормальным в случайном лесе из-за аппроксимации обучающей выборки. Модель стабильна, значение ниже всего на тесте.

3) Градиентный бустинг: предположительно в теории одна из лучших моделей, выявляет незаметные паттерны и проблемы, работает последовательно с большим числом деревьев. Наиболее хороша при наличии как у нас не только линейных связей. Показала наилучший результат из всех трех моделей по двум параметрам оценки, наименьшее значение RMSE также на обучающей выборке, как и MAPE в процентах. Значения теста и кросс-валидации для квадратичной ошибки примерно схожи, но чуть хуже CV на случайном лесе. Однако все такая же высокая точность прогноза.

Выводы: RMSE = 22 млн долларов говорит о том, что модель в среднем ошибается на 22 миллионов пи прогнозировании кассовых сборов. MAPE = 6.5% - средняя ошибка составляет чуть более 6.5% от истинных кассовых сборов. Относительно среднего значения нашей целевой переменной, результаты достаточно близки через оценку разными показателями и разными методами. MAPE показывает умеренную точность, так как меньше 10% ошибка с учетом нестабильности сборов фильмов в последние года. Наилучшим методом стал градиентный бустинг, наиболее точно предсказал на обучении (RMSE = 22.07, MAPE = 6.48%), случайный лес дал наилучший результат на тесте MAPE = 6.69%. Градиентный Бустинг максимизирует точность. МНК подходит для интерпретации более простых моделей, однако процентные оценки здесь примерно схожи во всех трех моделях, а RMSE показал лучше всего в МНК.

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 с исходными и подобранными значениями гиперпараметров.

значение RMSE на тестовой выборке с исходными и подобранными значениями гиперпараметров.

Проинтерпретируйте полученные результаты.

Сравнительный анализ по моделям:

1) Линейная регрессия оценивает коэффициенты, минимизируя квадраты ошибок. Соответственно, нет гиперпараметров, которые можем использовать для тюнинга, поэтому прибегнем к Лассо и Ридж Регрессии. Изначально результаты на кросс-валидации были ниже, чем на тестовой выборке, что говорит о понятности и четкой

применимости МНК оценки в данном случае при линейной связи, однако сильных изменений нет.

Ридж к минимизации суммы квадратов добавляет L2 регуляризацию в виде $a \cdot$ СУММ b_j^2, что решает проблему мультиколлинеараности и нестабильности оценок. Сила регуляризации от 1 перешла к 10, значения чуть ниже, чем в МНК. Модель устойчива к переобучению. Лассо добавляет вместо квадратов b_j , их модули, то есть L1 регуляризацию для отбора признаков, так как некоторые коэффициенты становятся равными нулю, модель становится проще для анализа при большом числе признаков. Здесь альфа с 1 наоборот уменьшилось до 0.01, то есть слабая сила практически не поменяла результат и нет избыточных признаков.

2) Случайный лес учитывает нелинейные связи, перекрестные переменные и не чувствительный к выбросам. Используем несколько разных гиперпараметров, значения Tuned логически становятся ниже после добавления тюнинга в случае тестовой модели и кросс-валидации. Вероятно, базовая модель переобучена, однако после включения гиперпараметров ошибка стала выше, что говорит об отсутствии пользы изменения глубины, листьев и разбиений (повышени с 22.91 до 24.9 в кросс-валидации)

3) Градиентный бустинг также показал хорошую точность, привел к снижению RMSE на несколько сотых. Стала лучше после уменьшения learning rate = 0.05 с 23.05 до 22.9. Однако показала чуть лучше результаты, чем случайный лес.

Результаты: градиентный бустинг оказался наиболее точным и устойчивым решением. Это указывает на то, что зависимость между характеристиками фильма и его коммерческим успехом носит сложный, но всё же системный характер, который лучше улавливается нелинейными ансамблевыми моделями. Линейные модели хорошо работают при ограниченном числе признаков и чёткой линейной зависимости, однако в условиях сложных взаимодействий между параметрами (между жанром и бюджетом) они становятся недостаточными. Тем не менее, они ценны как интерпретируемые базовые модели и показали, что признаки в данных подобраны достаточно грамотно. Поэтому результаты МНК оценки на долю сотых ниже, однако практически не изменились при проверке на тестовой выборке и по другим параметрам уступают ГБ.

Наилучший компромисс между точностью и устойчивостью в данной задаче обеспечивает градиентный бустинг, особенно при аккуратной настройке параметров. Эта модель может использоваться для прогнозирования сборов новых фильмов и поддержки принятия решений в области маркетинга и прокатной стратегии, так как

например тюнинг гиперпараметров дал прирост значения ошибки в Случайном лесе, что говорит об ухудшении модели, а не привычном улучшении.

	Model	Initial Params	Tuned Params	CV RMSE	Test RMSE
0	Linear Regression (OLS)	-	-	22.852050	22.866770
1	Ridge Regression	alpha=1.0	{'ridge__alpha': 10.0}	22.852023	22.866625
2	Lasso Regression	alpha=1.0	{'lasso__alpha': 0.01}	22.852045	22.866500
3	Random Forest (Base)	n_estimators=100	-	24.912775	25.250196
4	Random Forest (Tuned)	various	{'max_depth': 5, 'min_samples_leaf': 1, 'min_s...	22.906592	22.929742
5	Gradient Boosting (Base)	n_estimators=100, learning_rate=0.1	-	23.050076	22.985831
6	Gradient Boosting (Tuned)	various	{'learning_rate': 0.05, 'max_depth': 3, 'min_s...	22.930838	22.910339

Табл. 18. Значения гиперпараметров и RMSE на разных методах.

Повышенная сложность: 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.

Цель: сравнить для подхода к тюнингу гиперпараметров для модели ГБ. OOB-ошибка применяется, когда модель обучается на подвыборках, а Кросс-валидация является способом оценки обобщающей способности модели на разных разбиениях.

Для OOB мы получили 200 деревьев (бустов), более глубокие деревья = 3, что говорит о более сложных взаимодействиях модели, деление узлов выполняется, когда в узле 3 или более объектов (см. Табл.19) и что в каждом листе будет не менее 1 объекта, что повышает гибкость модели. Результаты показали для тестовой выборки RMSE на 3 десятых выше: 22.94, что хуже, чем у кросс-валидации. Это может быть связано из-за глубины деревьев, которые выдают низкую ошибку, но переносят меньшую обучающую способность на тестовую выборку.

На кросс-валидации обучение достаточно медленное = 0.05, но благодаря этому более устойчивое для бустинга. Деревья мельче = 3, а общее число деревьев ниже = 100, что предотвращает переобучения. На обучающей выборке RMSE = 22.93 примерно равно результату OOB, и тестовой выборке показало наиболее лучший результат среди двух подходов 22.91. Это говорит о том, что модель показывает лучшую обучающую способность, даже если она проще, что предпочтительно при проведении эконометрического анализа с сохранением точности и минимизации переобучения.

Результаты: разница в 0.03 на тестовой выборке между кросс-валидацией и ошибкой (см. Табл.19.) эквивалентна 30 тысячам долларов кассовых сборов, небольшая, но значимая сумма при моделировании и прогнозе сборов. В среднем ошибка, как и в пунктах выше, RMSE = 23, то есть цена ошибки прогноза 23 млн долларов.

	Параметры	**RMSE train**	**RMSE test**
OOB	{'n_estimators': 200, 'max_depth': 3, 'min_samples_split': 2, 'min_samples_leaf': 2}	**22.95**	**22.94**
CV	{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100}	**22.93**	**22.91**

Табл. 19. Лучшие результаты по кросс-валидации и OOB ошибки

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

В сфере киноиндустрии прогноз кассовых сборов — задача с высокой долей неопределённости, зависящая от множества факторов: жанра, бюджета, актёрского состава, даты релиза, маркетинга и даже общественных трендов. При выборе наилучшей модели для предсказания важно учитывать не только точность на обучающих и тестовых данных, но и устойчивость на новых выборках (кросс-валидации), а также интерпретируемость, стабильность и практическую применимость. Основные метрики, на которые мы ориентируемся: RMSE в миллионах долларов, MAPE — показывает, насколько в среднем предсказание отклоняется в процентах от реального значения, что особенно важно для понимания относительной точности модели при работе с фильмами разных масштабов (от малобюджетных до блокбастеров). Сравнение проводится по обучающей, тестовой и кросс-валидационной метрикам — чтобы оценить как переобучение, так и обобщающую способность моделей.

Лучшая модель: Градиентный Бустинг

На обучающей и тестовой выборке градиентный бустинг показал наименьшие ошибки в абсолютных (RMSE) и относительных (MAPE) показателях. Обучающий RMSE (22.07), что говорит о стабильности и отсутствии переобучения. Глубина деревьев

= 3 и learning rate = 0.05 обеспечивают хороший баланс между точностью и обобщающей способностью. Несмотря на то, что МНК оценка показала лучший RMSE на кросс-валидации (22.87), он хуже на тесте MAPE (6.48% напротив 6.67%), что делает его менее надёжным для реального прогноза кассовых сборов. Использование кросс-валидации для настройки параметров обеспечило лучшую переносимость модели на новые данные.

При RMSE = 22.07 модель в среднем ошибается на $22.07 млн, что меньше, чем у других моделей. MAPE = 6.48% — в среднем ошибка составляет всего 6.48% от кассовых сборов. Это соответствует хорошему уровню точности в прикладной эконометрике (обычно целевой уровень ≤10%). Он наиболее адаптивен к реальным рыночным закономерностям, где поведение зрителей и коммерческий успех фильмов зависят от совокупности факторов, а не от одного признака.

Худшая модель: Случайный лес

Несмотря на приемлемые значения ошибок линейной регрессии, случайный лес уступает по точности двум способам, так как не может уловить сложные нелинейные связи, которые явно присутствуют в зависимости между рейтингом, бюджетом, сиквелом и сборами. Его квадратичная ошибка выше других на 2 целых, то есть 24.91 на кросс-валидации и 25.25 на тесте, что сильно высоко при прочих равных. Мы не учитываем перекрестные и квадратичные взаимодействия переменных, а также эффекты структурных сдвигов (например, жанр, год, сезонность). СЛ высокая ошибка на кросс-валидации указывает на слабую обобщающую способность и уязвимость к переобучению в данном наборе данных.

Выводы: анализ показывает, что все три модели продемонстрировали достаточно близкие результаты, особенно по процентным ошибкам (MAPE — от 6.48% до 6.73%). Это говорит о том, что признаки в данных были хорошо подобраны, а уровень шума — умеренный. Однако, при более детальном рассмотрении становится заметна разница в поведении моделей: градиентный бустинг точнее моделирует нелинейности, которые типичны для коммерческого успеха фильмов. Его ошибки ниже как в абсолютном (RMSE), так и относительном (MAPE) значении, а это важно в экономике развлечений, где каждый процент точности влияет на решения о маркетинге, дистрибуции и бюджетировании. Оценка через CV и корректно подобранные гиперпараметры обеспечивают устойчивость прогноза, а не только точность на обучающей выборке.

Изначально МНК показывал чуть лучше результаты, он более легко и точно интерпретируется МНК в отличии от ГБ. ГБ устойчив к переобучению, также как и

МНК. МНК учитывает только линейные эффекты и менее гибкая, чем градиентный бустинг при анализе кинопроката.

Результаты: в условиях современной киноиндустрии, где бюджеты фильмов достигают сотен миллионов долларов, а коммерческий успех измеряется кассовыми сборами, точное прогнозирование выручки — ключ к эффективному управлению рисками, инвестициями и стратегией продвижения. МНК, хоть и прост в интерпретации и широко применим, ограничен своей линейной природой. В реальности кассовые сборы — это результат сложного взаимодействия факторов: нелинейного роста прибыли при превышении определённого бюджета (эффект "блокбастера"), усиленного влияния рейтинга при высоком уровне ожиданий, синергетических эффектов от сочетания известных актёров, режиссёра и бренда (франшиза). Такие связи невозможно корректно описать линейной моделью.

Градиентный бустинг способен выявлять и обучаться на сложных, многомерных и нелинейных закономерностях, благодаря ансамблю решающих деревьев. Он адаптируется под структуру данных, учитывает взаимодействия между переменными (например, эффект "дорогой фильм + высокий рейтинг"), устойчив к шуму и выбросам (всплески хайпа, эффект вирусности). Например, продюсерская компания использует модель для оценки потенциальных сборов будущего фильма. Если она применяет МНК, то получит линейную зависимость, где каждый доллар бюджета "равноценен", а рейтинг "влияет" одинаково на любой проект. Это упрощённая реальность, которая может привести к недооценке рисков или переоценке возможностей. Если используется градиентный бустинг, модель обучится, например, на том, что рейтинг особенно важен для малобюджетных фильмов, или что эффект сиквела усиливается при определённом бюджете и дате выхода. Такая модель даст точнее прогноз, что позволит скорректировать бюджет, пересмотреть рекламную кампанию или изменить дату релиза.

4.5 Повышенная сложность: включите в анализ дополнительный метод регрессии, 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.

Huber Loss - функция потерь, используемая в регрессионных задачах, которая объединяет преимущества MSE и MAE. Наша задача ее минимизировать при

соответствующих параметрах. Сначала Huber loss вычисляет разницу между фактическими и предсказанными значениями. Если эта разница меньше порогового значения (дельты), то функция ведёт себя как MAE, в противном случае переключается на MSE. Применяет поведение как MSE при малых ошибках, высокая чувствительность, гладкость градиента, хорошее поведение при обучении. Поведение как MAE при больших отклонениях: делает функцию устойчивой к выбросам. Поэтому она является более устойчивой к выбросам, чем чистый MSE, и при этом сохранять плавность градиента, в отличие от MAE, у которой градиент может быть разрывным.

Преимущества функции потерь:

- Дифференцируемость: Huber Loss непрерывна и дифференцируема во всех точках, включая переход между квадратичной и линейной частью. Это важно для корректной работы градиентных методов оптимизации.
- Устойчивость к выбросам: в отличие от MSE, который сильно штрафует большие ошибки, Huber Loss переходит от квадратичного поведения к линейному при превышении порога. Это снижает влияние выбросов и делает модель более устойчивой к шуму.
- Плавный ландшафт оптимизации: плавный переход между режимами (MSE и MAE) обеспечивает стабильность градиентов, снижая риск взрыва или исчезновения градиента.
- Гибкость через параметр δ: параметр δ позволяет настраивать чувствительность к ошибкам — можно адаптировать функцию под конкретную задачу.

Недостатки:

- Необходимость настройки параметра. Порог δ влияет на поведение функции и требует тщательной настройки, часто через кросс-валидацию.
- Не всегда лучшая для всех задач: несмотря на устойчивость к выбросам, Huber Loss может уступать другим функциям потерь, если задача требует высокой точности при малых ошибках.
- Слабее на малых ошибках, чем MAE: меньше штрафует за небольшие ошибки по сравнению с MAE, что может немного снизить точность в задачах без шума.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

В задачах прогнозирования кассовых сборов фильмов, где есть выбросы (редкие блокбастеры с аномально высокими сборами), большинство фильмов укладываются в

диапазон обычных значений (скажем, до $100 млн), делаем так, чтобы пара выбросов не «перетягивали» модель, как это бывает с MSE. HL делает модель устойчивой к редким пиковым фильмам, сохраняя точность в массе предсказаний.

При интерпретации логов значение Huber Loss постепенно уменьшается, что говорит о стабильном обучении модели. Метрики на тестовой выборке: MSE: 16680, Huber Loss: 126.24, RMSE: 129.15 гораздо ниже начального значения, что указывает на хорошую устойчивость модели.

Сопоставление: модель, обученная с использованием Huber Loss, показала устойчивость к выбросам, что может быть полезно, если цель — снижение влияния отдельных «аномальных» фильмов. Однако, для точности предсказаний в широком диапазоне кассовых сборов, градиентный бустинг остаётся наиболее эффективным решением.

RMSE у модели с Huber Loss существенно выше: 129.15 против 22.07 у ГБ, то есть модель переобучилась или недообучилась, и не обеспечила высокую точность на тестовой выборке. Это может быть из-за не сложной функциональной формы, малом количестве признаков. Градиентный бустинг показал лучшую общую производительность, особенно по RMSE и MSE (см. Табл. 20)

Метрика	HL модель	Градиентный бустинг
RMSE	129.15	22.07
MSE	16680	743
Huber Loss	126.24	-

Табл. 20. Сравнение модели Huber Loss с лучшей из предыдущего анализа

4.6. В случае необходимости проведите дополнительный анализ.

Опираясь на сравнительный анализ из пунктов 4.1-4.5, была выявлена лучшая и худшая модель при сравнении на разных метриках и разными способами, в том числе используя кросс-валидацию. Результаты релевантные и обоснованные, которые не требуют дополнительных проверок или включения иных моделей, так как достаточно полно раскрывают суть построения регрессий на кассовые сборы и интерпретации соответствующих признаков.

5. Эффекты воздействия

При выполнении данного задания необходимо объединить обучающую и тестовую выборки в одну.

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

В настоящем исследовании для целевой переменной имеется 2 исхода:

$$Y_i = \begin{cases} Y_{1i}, & \text{если } D_i = 1 \\ Y_{0i}, & \text{если } D_i = 0 \end{cases} = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

Итак, целевая переменная (кассовые сборы) потенциально может существовать в двух ситуациях: наличие у фильма возрастного рейтинга R ($D_i = 1$) и его отсутствие ($D_i = 0$). Для каждого фильма мы может наблюдать только 1 из них. Так, ожидается, что в случае наличия у фильма возрастных ограничений кассовые сборы будут ниже, чем в его отсутствии.

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

Перед оценкой эффектов необходимо выделить группу соблюдателей, то есть режиссеров, который будут снимать фильм с рейтингом R, если до этого они снимали только такие фильмы и не будут этого делать, если их фильмография более разнообразна. Для этого введем некую случайную величину $U_i \sim U[0,1]$ и введем гипотетические переменные:

$$r_{1i} = I(P(D_i = 1 |\, X_i, Z_i = 1) \geq U_i)$$

$$r_{0i} = I(P(D_i = 1 | \mathrm{X}_i, Z_i = 0) \geq U_i)$$

К соблюдателям будем относить тех, у кого $r_{1i} > r_{0i}$.

Для начала было необходимо оценить настоящий эффект воздействия на симулированных исходах при разных ситуациях, где $TE_i = Y_{1i} - Y_{0i}$. Средний эффект воздействия был посчитан как $ATE = E(Y_{1i} - Y_{0i})$. Для расчета локального среднего эффекта воздействия была использована следующая формула:
L$ATE = E(Y_{1i} - Y_{0i}|r_{1i} > r_{0i})$, для расчета условного среднего эффекта воздействия использовалась следующая формула: $CATE_i = E(Y_{1i}|X_i) - E(Y_{0i}|X_i)$.

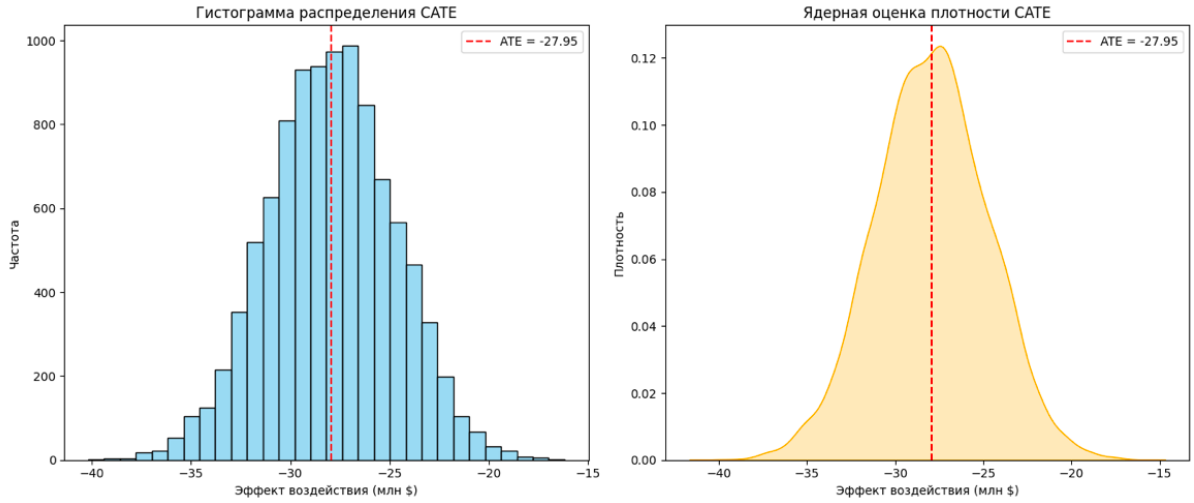
Исходя из результатов оценивания (см. Табл 21) можно сделать вывод о том, что в среднем получение рейтинга 18+ снижает кассовые сборы на 12,83 миллиона долларов, при прочих равных. Результат были предсказуемым, так как согласно нашей предпосылки сборы были замоделированы именно таким образом. Локальный средний эффект свидетельствует о том, что для последователей негативный эффект меньше. Возможно, инструментальная переменная связана с факторами, которые смягчают негативный эффект от взрослых сцен, например, лояльная аудитория режиссера.

Во время оценивания локальных эффектов воздействия средний эффект воздействия оказался равен -27.95, что указывает на значительное снижение кассовых сборов для фильмов с рейтингом 18+. Гистограмма распределения CATE имеет околонормальное распределение. Ядерная оценка плотности имеет также околонормальное распределение с двумя локальными пиками (см Рис 2).

	Метрика	Значение
0	ATE (Средний эффект)	-12.834587
1	LATE (Локальный средний эффект)	-10.654991

Таблица 21. Сравнительная таблица для среднего и локального среднего эффектов воздействия

Рис 2. Гистограмма распределения CATE и ее ядерная оценка плотности



5.3 Оцените средний эффект воздействия как разницу в средних по выборкам тех, кто получил и не получил воздействие. Опишите недостатки соответствующего подхода с учетом специфики рассматриваемой вами экономической проблемы. Примечание: в этом пункте и далее, если не сказано иное, используются лишь наблюдаемые значения целевой переменной.

Сделаем допущение о независимости: $E(Y_{1i}|D_i = 1) = E(Y_{1i})$ и $E(Y_{0i}|D_i = 0) = E(Y_{0i})$.

После оценки получилось, что наличие сцен 18+ повышает кассовые сборы на 1,26 миллиона долларов, что не является правдой по построению выборки. К недостаткам метода относится игнорирование проблемы эндогенности. Так, получение рейтинга R не является случайным процессом, решение о присвоении возрастного рейтинга принимается комиссией на основании многих характеристик продукта. Также возможна обратная причинно-следственная связь: фильмы, которые стремятся стать высокобюджетными, сознательно будут избегать присвоения рейтинга 18+, а низкобюджетные картины могут стремиться к поднятию охватов за счет провокационного контента. Кроме того, наивный предиктор ошибся в знаке коэффициента сильно завысив эффект от наличия сцен 18+.

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

- **метода наименьших квадратов.**
- **условных математических ожиданий.**
- **взвешивания на обратные вероятности (в случае возникновения ошибок убедитесь в отсутствии оценок вероятностей, равных 0 или 1 и при необходимости измените метод оценивания).**
- **метода, обладающего двоичной 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. Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе, и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

В предыдущих пунктах было получено, что среди классификационных моделей наилучшей является модель логистической регрессии, а среди регрессионных – модель градиентного бустинга.

Ослабим предыдущую предпосылку о независимости и введем предпосылку об условной независимости: $(Y_{1i}|D_i = 1, X_i) = E(Y_{1i}|X_i)$ и $E(Y_{0i}|D_i = 0, X_i) = E(Y_{0i}|X_i)$, тогда возможно оценить средний эффект воздействия с помощью нескольких методов. Предпосылка может нарушаться, если в модели присутствуют неучтенные смешивающие факторы. Например, бюджет маркетинговой кампании. Так, если фильм для взрослой аудитории будет хорошо прорекламирован, то возможно он соберет большую кассу, чем киноленты на широкую аудиторию, также при высокобюджетной маркетинговой стратегии режиссеры могут не обращать внимания на возрастные ограничения и оставлять в киноленте любые необходимые сцены. Также возможна проблема самоотбора, когда режиссеры нишевых проектов отдают предпочтения определенным сценам.

По результатам оценивания среднего эффекта воздействия с помощью метода наименьших квадратов было получено, что наличие у фильма рейтинга R приводит к уменьшению кассовых сборов на 1.05 миллиона долларов, что задает правильный знак

воздействия, но все еще далеко от истинного значения. Также было получено, что метод более предпочтителен, в сравнении с наивным. Однако, заметим, что данный метод дает сильно смещенную оценку из-за нелинейных связей между кассовыми сборами и остальными факторами.

Средний эффект воздействия при оценивании с помощью условных математических ожиданий показывает, что фильмы с возрастным рейтингом R будут зарабатывать в прокате на 13,3 миллиона меньше при прочих равных.

При подсчёте среднего эффекта с помощью взвешивания обратных вероятностей было получено, что наличие определенных сцен уменьшают кассовые сборы на 12,24 миллиона.

Использование метода двойной устойчивости позволило получить иные результаты. Так, возрастной рейтинг R уменьшает кассовые сборы на 11,2 миллионов.

При оценивании с помощью двойного машинного обучения было получено, что возрастной рейтинг R уменьшает кассовые сборы на 13 миллионов, что является довольно близким результатом к истинному.

По результатам оценивания, можно заключить, что наилучшим методом оценивания среднего эффекта воздействия оказался метод двойного машинного обучения. Различия с истинным значением для этого метода минимальны.

Для выполнения задания повышенной сложности был включен метод синтетического контроля. Метод создает синтетическую контрольную группу как взвешенную комбинацию единиц контрольной группы. Результаты показали, что наличие рейтинга 18+ приводит к потере 12 миллионов кассовых сборов, что является довольно близким результатом к истинному. К преимуществам метода можно отнести учет нелинейных зависимостей, устойчивость к искажению результатов из-за смешивания взаимосвязей с эффектами от других переменных, а также метод подходит для малого числа обработанных единиц. К недостаткам метода относится лучшая работа на агрегированных данных, а не на индивидуальных. Также метод чувствителен к выбору ковариат и имеет более сложную процедуру построения доверительных интервалов.

5.5 Оцените локальный средний 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.

Повышенная сложность: воспользуетесь также параметрической моделью, например, с помощью пакета switchSelection. Обсудите преимущества и недостатки такого подхода по сравненинию с двойным машинным обучением. Обычный метод инструментальных переменных параметрическим подходом не считается.

Представим, что переменная $budget_i$ отсутствует в данных, что приводит к эндогенности в данных, так как бюджет влияет как на кассовые сборы, так и может влиять на возрастной рейтинг кинокартины из-за желания продюсеров получить прибыль и привлечь как можно больше зрителей.

После оценки LATE с помощью двойного машинного обучения значение получилось 1,41, а с использованием инструментальной переменной -- -16,13, что близко к истинному значению -10,65.

Средний эффект воздействия показывает общее влияние для выборки, но также может усреднять разнонаправленные эффекты для разных подгрупп. Локальный средний эффект оценивает влияние только для compliers, то есть тех фильмов, которые меняют свой возрастной рейтинг при смене инструмента (в нашем случае, скорее смене режиссера). Разница в оценках может быть объяснена гетерогенностью эффектов. Например, для режиссеров, которые всегда выбирают рейтинг R эффект может снижаться из-за лояльной аудитории.

Значение локального среднего эффекта означает, что для последователей присвоение возрастного рейтинга R снижает кассовые сборы примерно на 16 миллионов.

В качестве дополнительной параметрической модели была выбрана двухэтапная модель с поправкой Хекмана. На первой стадии оценивается Пробит-модель, на основе той модели вычисляется поправка на смещение самоотбора. На второй стадии оценивается 2SLS с поправкой Хекмана, то есть влияние возрастного рейтинга на выручку с учетом контрольных переменных. Результат показал, что локальный эффект воздействия для этой модели составил -10,73, что довольно близко к истинному значению. Метод является наилучшим среди всех использованных.

5.6 Оцените условные средние эффекты воздействия с помощью:

• метода наименьших квадратов • S-learner.

• T-learner.

• метода трансформации классов. • X-learner.

Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при реализации государственных программ. Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

В настоящем исследовании применение метода X-learner может быть мотивировано несбалансированностью данных, так фильмов с иными возрастными рейтингами, отличными от R в выборке значительно меньше. Также метод способен учитывать неоднородность факторов, например неоднородность бюджета для кинокартин. При этом метод менее чувствителен к гетероскедастичности ошибок.

Сравнение результатов планируется в следующем пункте, однако, можно сказать, что методы предоставляют довольно разные и неоднородные результаты для каждого наблюдения.

Для каждого конкретного фильма был посчитан свой результат CATE. При создании фильма продюсерам стоит обращать внимание на значение данного показателя. Так, если CATE принимает положительное значение, то наличие у фильма рейтинга R положительно сказывается на кассовых сборах и при релизе стоит оставить все запланированные сцены, также рекомендуется проведение усиленной маркетинговой кампании среди взрослого населения. Напротив, при отрицательном значении CATE стоит попробовать получить более смягченный возрастной рейтинг, чтобы не выпускать убыточный проект, создателям стоит сделать акцент на семейную аудиторию.

В качестве альтернативного метода взят Casual Forest. Этот метод является расширением идеи Random Forest. Каждое дерево строится по принципу максимизации разницы в эффектах между подгруппами, при этом используется специальный критерий расщепления, учитывающий разницу средних исходов в узле и баланс ковариат между группами. Для каждого наблюдения эффект оценивается как средневзвешенное эффектов по деревьям, в которые он попал.

$$CATE_i = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{i\in L} T_i Y_i}{\sum_{i\in L} T_i} - \frac{\sum_{i\in L}(1-T_i)Y_i}{\sum_{i\in L}(1-T_i)}$$

Где L – лист дерева, который содержит X.

К преимуществам метода можно отнести автоматическую балансировку ковариат между группами, обнаружение сложных взаимодействий, а также устойчивость к переобучению. Недостатками метода является требование большой выборки, зависимость от гиперпараметров, а также сложность в интерпретации

5.7. Выберите лучшую модель оценивания условных средних эффектов воздействия, используя: • истинные значения условных средних эффектов воздействия. • прогнозную точность моделей. • псевдоисходы. Проинтерпретируйте различия в результатах различных подходов.

Используя истинные значения условных средних эффектов воздействия, оказалось, что наилучшим методом является X-learner, имея наименьший MSE. Наихудшим методом стал метод трансформации классов.

Основываясь на результатах равнения прогнозной точности моделей, можно заключить, что наилучшей является модель, построенная методом наименьших квадратов. Напротив, судя по псевдоисходам наилучшим является метод трансформации классов.

Заметим, что результаты методов совершенно разные. Это может происходить из-за разных целевых метрик при сравнении, так в одном случае минимизируется разница эффектов, а в другом разница в предсказании исхода. Также в модели могут быть разные допущения о данных, что влияет на итоговый выбор.

5.8 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. Сделаете вывод об устойчивости результатов к качеству используемых методов машинного обучения.

Худшими из моделей получились модель случайного леса и модель ближайших соседей KNN. Для этого воспользуемся методом двойного машинного обучения, как лучшим методом предыдущего пункта. Для поиска локального среднего эффекта наилучшей моделью оказалась параметрическая модель повышенной сложности, однако она не использовала классификационные и регрессионные модели, поэтому для этого

пункта была взята вторая лучшая – двойное машинное обучение с использованием инструментальной переменной.

Исходя из полученных оценок (см. Табл 22) можно получить, что эффекты воздействия оценились намного хуже, чем в предыдущих пунктах, из чего можно сделать вывод, о том что качество используемых методов машинного обучения играет важную роль.

	Оценка
ATE	-12.834587
ATE naive	1.259453
ATE ls	-1.052533
ATE_gb	-13.297501
ATE_IPW	-12.248003
ATE_DR	-11.133666
ATE dml standard	-17.766180
ATE_SC	-12.000887
ATE dml worst	-16.558831
LATE	-10.654991
LATE dml iv	-14.660768
LATE par. model	-10.730000
LATE_dml_iv_worst	-13.505978

Таблица 22. Результат оценивания средних эффектов воздействия и локальных эффектов воздействия разными способами

5.9 Резюмируете ключевые выводы проведенного в данном разделе анализа.

В данном разделе был проведен анализ методов, для оценки средних эффектов воздействия, локальных эффектов воздействия и условных эффектов воздействия на основе результатов анализа прошлых пунктов.

Обратим особое внимание на то, что при оценке эффектов воздействия на реальных данных особенно важно провести предварительный анализ методов регрессии и классификации для выявления наилучшего. В ином случае существует риск сильных расхождений истинных значений с посчитанными.

5.10 В случае необходимости проведите дополнительный анализ.

В качестве дополнения возможен анализ иных методов оценки, а также параметрических моделей, в частности: анализ чувствительности к инструментальной

переменной, проверка значений эффектов на подвыборках, а также сделать визуализацию результатов, разделив фильмы по жанрам.

Список литературы

1. Люди в черном // Кинопоиск URL: https://www.kinopoisk.ru/film/1091/?ysclid=mazpmo0zjm99928730&utm_referrer=yandex.ru (дата обращения: 22.05.2025).
2. После нашей эры // Кинопоиск URL: https://www.kinopoisk.ru/film/577285/?ysclid=mazpor0ovt398823342 (дата обращения: 22.05.2025).
3. Conaway, B., & Ellis, D. (2015, январь). Do MPAA ratings affect box office revenues? Academy of Business Research Journal, 1, 64–88
4. De Vany, A., & Walls, W. D. (2000). Does Hollywood make too many R-rated movies? Risk, stochastic dominance, and the illusion of expectation. The Journal of Business, 75(3), 425–451. https://doi.org/10.1086/338705
5. Greenaway, M. & Zetterberg B., 2012. Success in the Film Industry: What Elements Really Matter in Determining Box-Office Receipts
6. Lazear, E. P. (2004). The Peter Principle: A theory of decline. Journal of Political Economy, 112(S1), S141–S163
7. Topf, P. (2010). Examining success in the motion picture industry. The Park Place Economist, 18(15), 1–8. (Т. 18, № 15, с. 1)
8. Huber Loss – Loss function to use in Regression when dealing with Outliers. (2023). MLexplained blog.