Национальный исследовательский университет
Высшая Школа Экономики, Санкт-Петербург

Кассовые сборы фильмов: Что на них влияет?

Дасмаева Мария, Рахманова Амина,
Клечикова Маргарита, Жигульская Евгения, Федоров Дмитрий

Дата защиты: 18 марта, 2021

Дата выполнения ..... 1-18 Марта, 2021
Дисциплина ..... Прикладная Статистика
Преподаватели ..... Холодилин Константин Аркадьевич
..... Полякова Евгения Юрьевна
..... Щапов Дмитрий Сергеевич

Содержание

1 Постановка вопроса

В данной главе мы рассказываем, что изучаем и зачем.

Наше исследование, основанное на анализе факторов, сопутствующих производству и выпуску фильма в прокат, посвящено оценке потенциальных кассовых сборов.

Прогнозирование кассовых сборов стало неотъемлемой частью киноиндустрии и потребностью как для продюсеров, так и для кинопрокатчиков. Во-первых, необходимо, чтобы кассовые сборы покрывали все экономические издержки и возмещали производственный бюджет киноленты. Вторая причина – формирование объемов рекламной кампании, которая нацелена на пробуждение заинтересованности у зрителей. Прогнозирование также стало важным условием и для достижения наибольшей экономической выгоды кинотеатров. Предпочтения отдаются тем фильмам, что способны продемонстрировать максимальную заполняемость кинозалов в пиковые периоды и поддерживать спрос со стороны зрителей в течение длительного периода проката.

Исходя из этого, в киноиндустрии были определены некоторые показатели, по которым можно определить, будет ли фильм успешным и какую прибыль он может принести как для кинодистрибьютора, так и для киносетей. К этим показателям относятся: дата выхода фильма (месяц, день недели), жанр, возрастные ограничения, актеры и даже культурные или религиозные нормы определенной страны проката. Совокупность правильно подобранных факторов обеспечит фильму высокий рейтинг и максимальные кассовые сборы.

Собственно, в этом и заключается наше исследование: найти те самые факторы, которые могут повлиять на кассовые сборы фильмов.

2 Обзор литературы и выдвижение гипотезы

В данной главе мы оцениваем наш вклад в существующую литературу и обсуждаем, что дополняем своим исследованием.

2.1 Обзор литературы

В нашей работе были использованы 2 датасета, взятые с сайта *www.imdb.com*

1. **IMDb_movies** – датасет с фильмами, включающий в себя 85855 наблюдений по 22 переменным.

2. **IMDb_ratings** – датасет с рейтингами фильмов, включающий в себя 85855 наблюдений по 49 переменным.

2.2 Выдвижение гипотезы

Гипотеза исследования:

Существуют факторы, влияющие на кассовые сборы фильмов, которые могут быть определены с помощью предиктивных моделей и механизмов программирования.

2.3 Наш вклад в существующую литературу

В ходе данного исследования мы найдем некоторые закономерности в киноиндустрии, которые тем или иным образом влияют на кассовые сборы фильмов. Таким образом, получим модель, которая поможет описать алгоритм действий кинодистрибьютера и киносетей для получения наибольшей выгоды от проката. Эта модель может значительно облегчить понимание важности тех или иных факторов при съемке и прокате картин.

3 Описание данных

3.1 Работа с переменными

В данной главе мы рассказываем про достоинства, недостатки и свойства наших данных и обсуждаем, почему они релевантны для нашего исследования.

В качестве преобразования данных мы соединили две таблицы в один датасет **imdb**, перевели номинальные переменные в категориальный формат, а количественные - в бинарный. Затем отфильтровали пропущенные значения в переменной с кассовыми сборами и убрали те наблюдения, в которых эта информация была приведена не в долларах. Далее разделили дату выпуска на месяц и день, а также добавили дни недели. После этого отфильтровали датасет, оставив только наиболее важные переменные, которые могут влиять на сборы. Таблица переменных, используемых нами в исследовании, вместе с их расшифровкой приведена далее.

Список используемых переменных

Переменная	**Расшифровка**
year	год выпуска фильма
month	месяц выпуска фильма
weekday	день недели выпуска фильма
genre	жанр фильма
duration	длительность (в минутах)
country	страна производства
language	язык(и) фильма
avg_vote	средний рейтинг пользователей
votes	количество голосов от пользователей
budget_currency	валюта бюджета
budget	бюджет фильма
usa_gross_income	кассовые сборы в США
worldwide_gross_income	кассовые сборы по всему миру
metascore	рейтинг Metascore (от 0 до 100)
reviews_from_users	количество отзывов от пользователей
reviews_from_critics	количество отзывов от критиков
total_votes	общее количество отзывов
weighted_average_rating	средневзвешенный рейтинг
median_vote	медианный рейтинг
males_allages_avg_vote	средний рейтинг от мужчин
males_allages_votes	среднее количество голосов от мужчин
females_allages_avg_vote	средний рейтинг от женщин
femles_allages_votes	среднее количество голосов от женщин
us_voters_rating	рейтинг от пользователей из США
us_voters_votes	количество голосов от пользователей из США
non_us_voters_rating	рейтинг от пользователей не из США
non_us_voters_votes	количество голосов от пользователей не из США

3.2 Достоинства и недостатки данных

Большинство данных наглядно отражают, насколько был успешен тот или иной фильм. Исходя из этого, можно определить, какие факторы стали тому причиной. Для рассмотрения этого вопроса можно взять такие переменные, как: кассовые сборы по всему миру, кассовые сборы в США, средний рейтинг пользователей, общее количество отзывов и другие

Говоря о недостатках, необходимо отметить, что некоторые данные отсутствовали в исследуемых датасетах, что создавало погрешности при вычислениях (например, пропущенные значения в переменной с кассовыми сборами). Кроме этого, сборы не всегда исчислялись в долларах (например, индийские рупии), что усложнило проведение исследования.

3.3 Обоснование выбора данных

Предоставленные данные наилучшим образом формируют основу для дальнейших заключений, так как наилучшим образом отражают все взаимосвязи между социально-экономической успешностью фильма и сторонними факторами, влияющими на уровень кассовых сборов (например: бюджет фильма, жанр, год выпуска, страна производства и др.). Кроме того, следует отметить, что в нашем исследовании мы используем большое количество переменных, что позволяет получить наиболее полную картину при анализе данных.

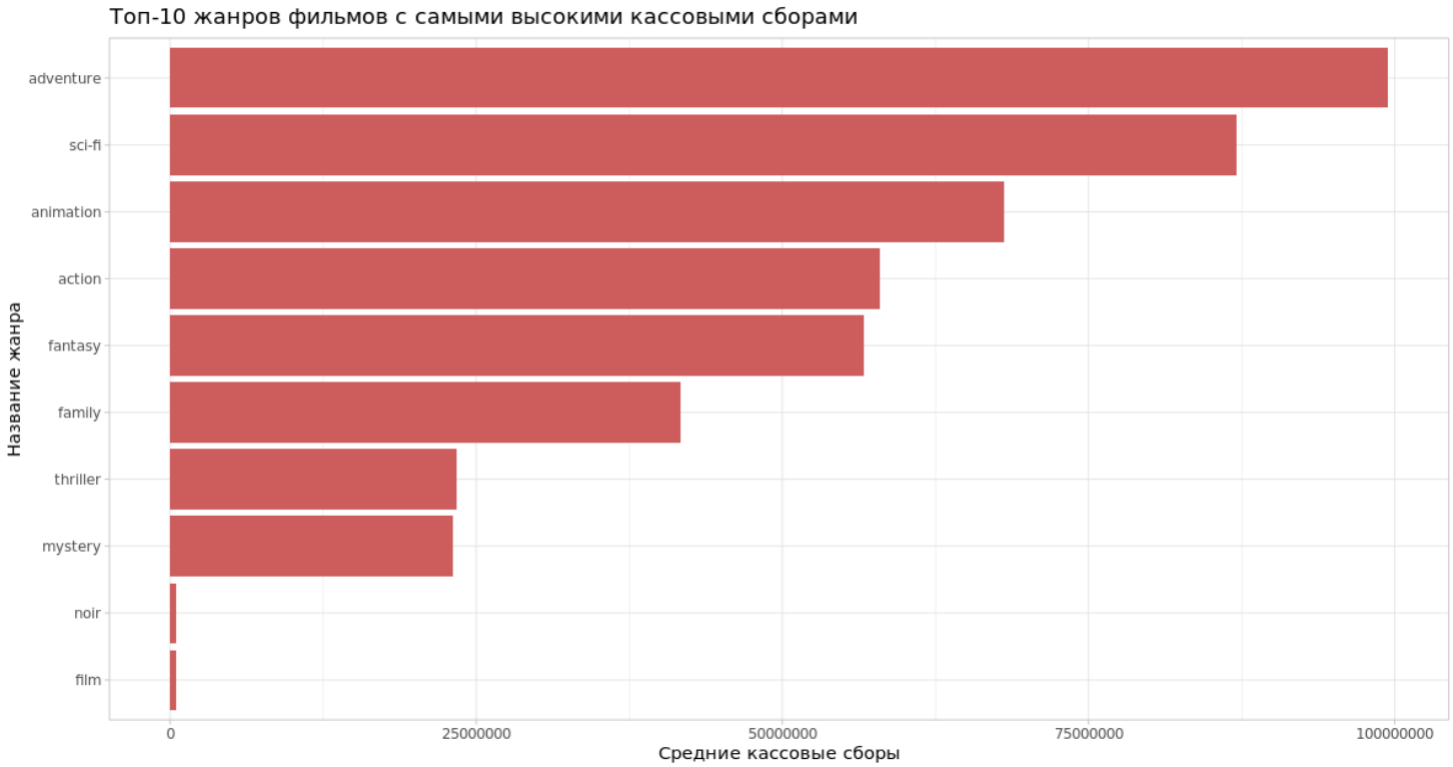
4 Методы исследования

В данной главе мы рассказываем, как из наших данных извлекаем ответ на исследовательский вопрос и какие методы для этого используем.

4.1 Методы

В нашей проектной работе мы использовали несколько методов исследования. В первую очередь мы рассмотрели несколько столбчатых диаграмм, а именно «10 жанров с самыми высокими кассовыми сборами» и «зависимость сборов от даты выхода фильма» (дня недели, месяца и года). Кроме того, мы провели статистический анализ, рассчитали коэффициенты Пирсона и Спирмена, а также построили предиктивную модель, которая будет описана далее.

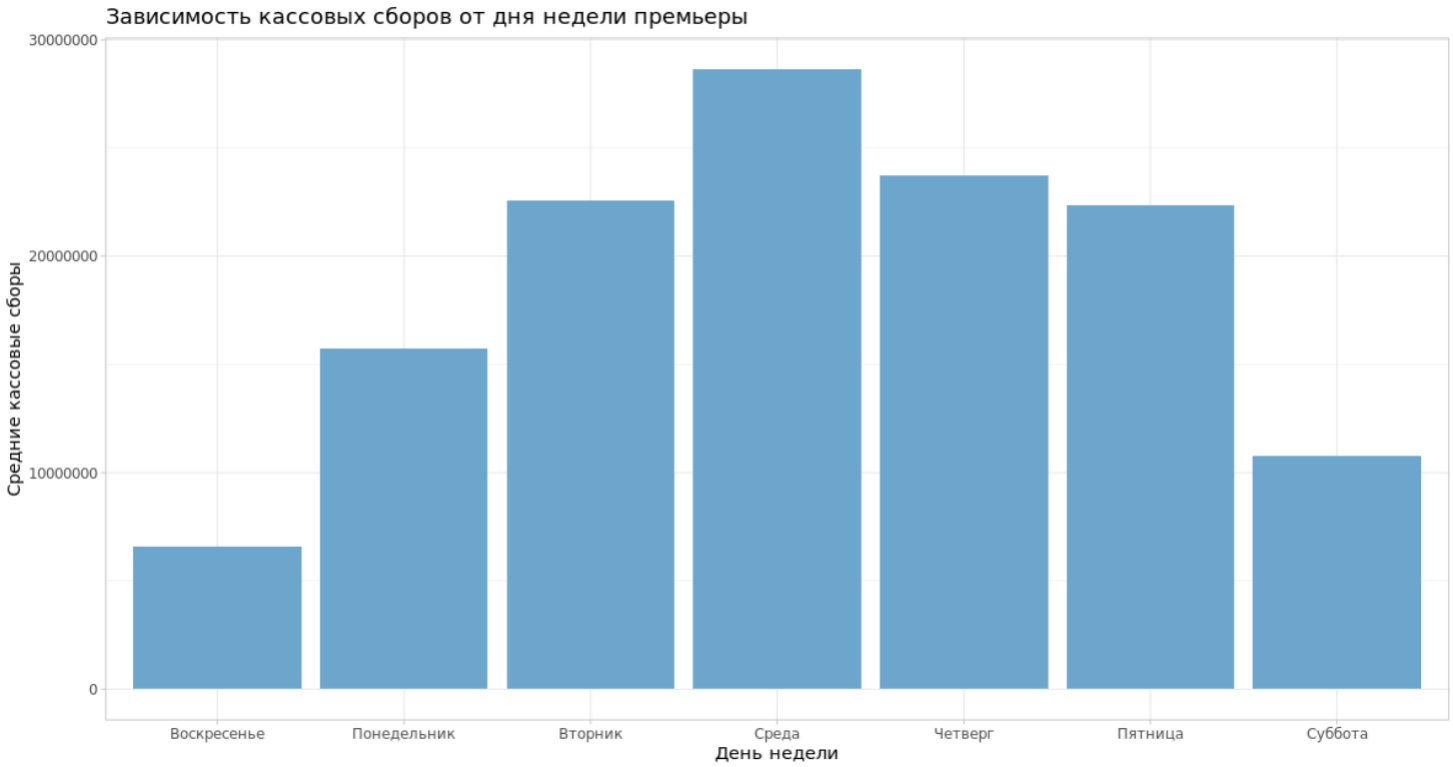
4.2 Графики



10 жанров с самыми высокими кассовыми сборами

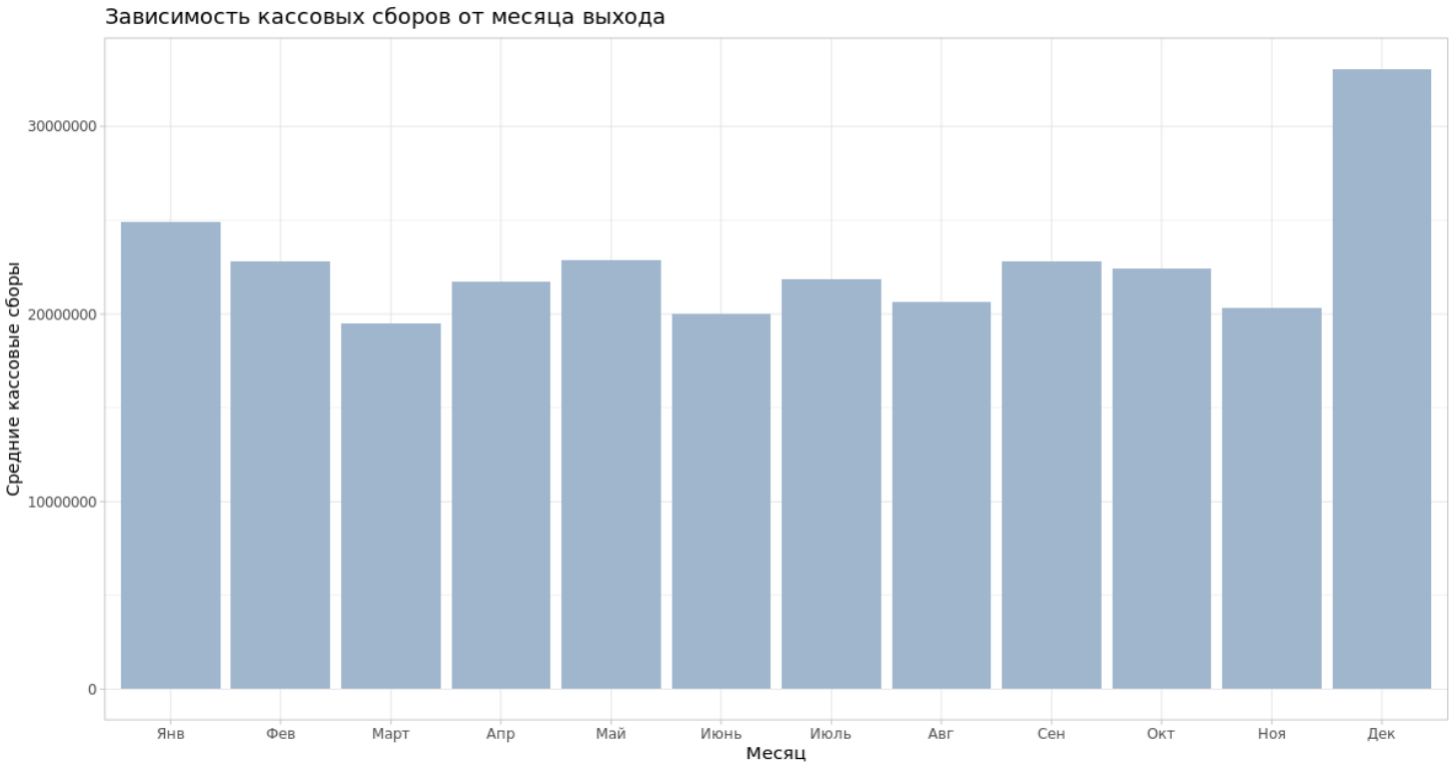
Приключенческие фильмы, исходя из графика, являются самыми кассовыми в киноиндустрии. Заинтересованность публики можно объяснить доступностью и простотой таких кинолент. Подобные фильмы не отличаются запутанным сюжетом и будут поняты зрителям всех возрастов. Научная фантастика, занимающая второе место, привлекает своей неординарностью и нереалистичностью, зацепляет зрителя новой идеей.

Далее мы решили проверить, зависят ли кассовые сборы от дня недели/месяца/года выхода фильма.



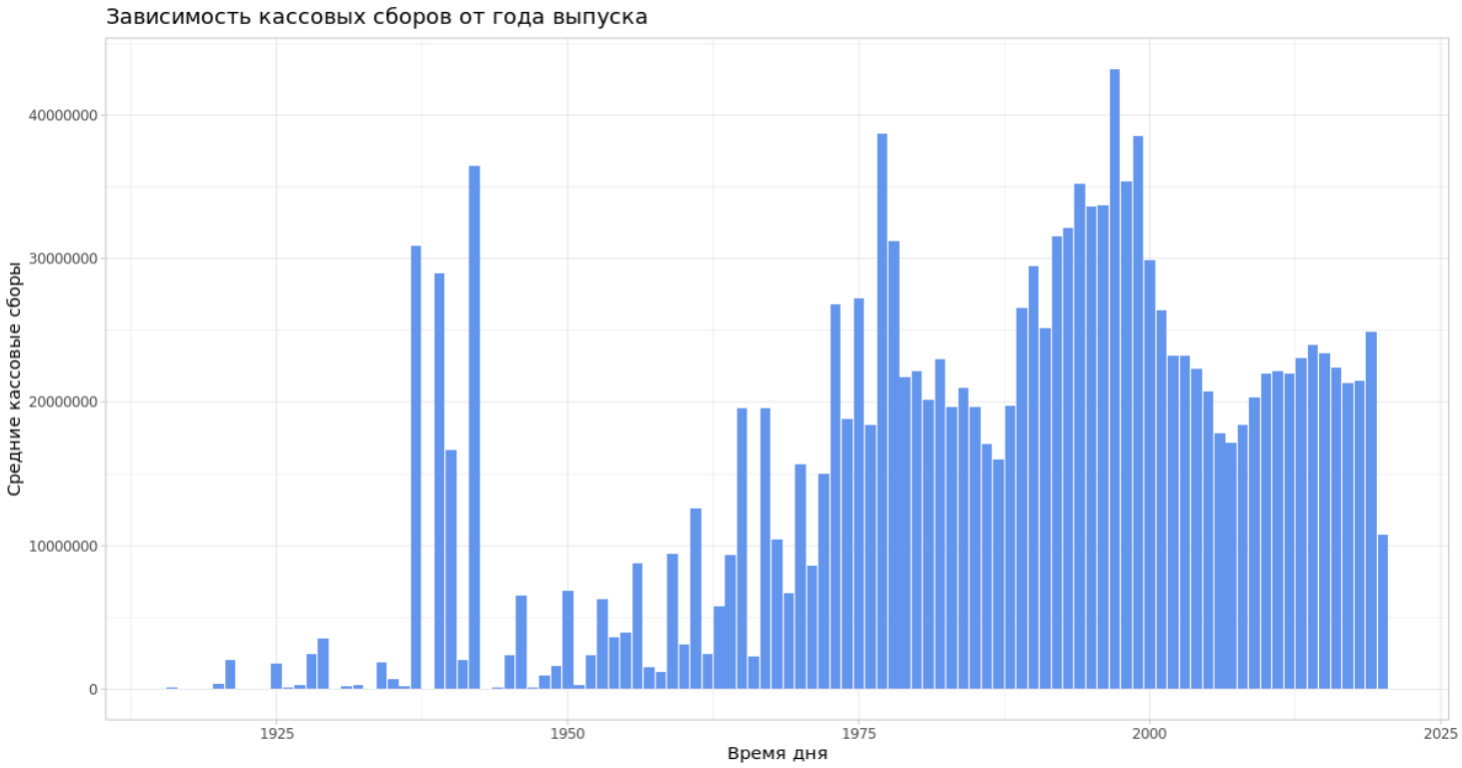
Зависимость кассовых сборов от дня недели

Выход большинства мировых премьер не опирается на конкретный день недели, а отталкивается от каких-то внешних факторов. Судя по графику, среда является наиболее «продаваемым» днем недели в отношении выхода фильма. И этот факт легко объяснить. Выход фильмов в середине недели позволяет им получить достаточную огласку со стороны критиков, что в свою очередь подогреет интерес зрителей к уикенду проката. Таким образом, пик посещаемости выпадет на часть недели с пятницы по воскресенье. Именно по этому периоду киносети предоставляют статистику «старта» новинок. Кроме того, выход в середине недели позволяет удобно проводить репертуарное планирование в случае большого количества негативных отзывов или технического брака кинолент. Тем не менее, некоторые страны все же ориентируются на конкретные дни недели. Например, в России - это четверг, во Франции – среда, в США – пятница или среда.



Зависимость кассовых сборов от месяца

Опираясь на следующий график, можно заключить, что декабрь – лучшее время для выпуска фильма. Последний месяц года – тяжелый и одновременно праздничный (именно фильмы на рождественскую тематику обычно наиболее популярные в декабре). Эти факторы отлично подходят для популяризации кинопроката посредством активной рекламы и выхода новых картин - людям необходим отдых и положительные эмоции, а просмотр фильма в кинотеатре предоставит им и то, и другое. Отголоски новогодних праздников отражаются и в январе: киносети все еще испытывают повышенный спрос со стороны зрителей, но ситуация приходит к усредненному среднегодовому значению.



Зависимость кассовых сборов от года выпуска

Несложно заметить немногочисленные пиковые значения кассовых сборов в первой половине XX века. Это легко объясняется появлением новых технологий (первые цветные фильмы), которые значительно повысили интерес зрителей к кинотеатрам. Не трудно догадаться, что одно из этих значений приходится на «Унесенные ветром» Виктора Флеминга. Этот фильм до сих пор остается одним из наиболее кассовых даже в сравнении с современными картинами. В конце 80-х, 90-х и начале 2000-х мы видим новые пиковые значения. И это тоже легко объяснить: именно в этот период происходит глобальный рост популярности кинотеатров, количество точек киносетей увеличивается в невероятных масштабах, а студии в Голливуде снимают общепринятые блокбастеры и киношедевры, популярные и по сей день. Помимо этого, дистрибьютеры находят все новые и новые методы заработка: продакт-плейсмент, современные виды рекламы, оптимизированные периоды проката. Все это помогает им добиваться новых кассовых рекордов.

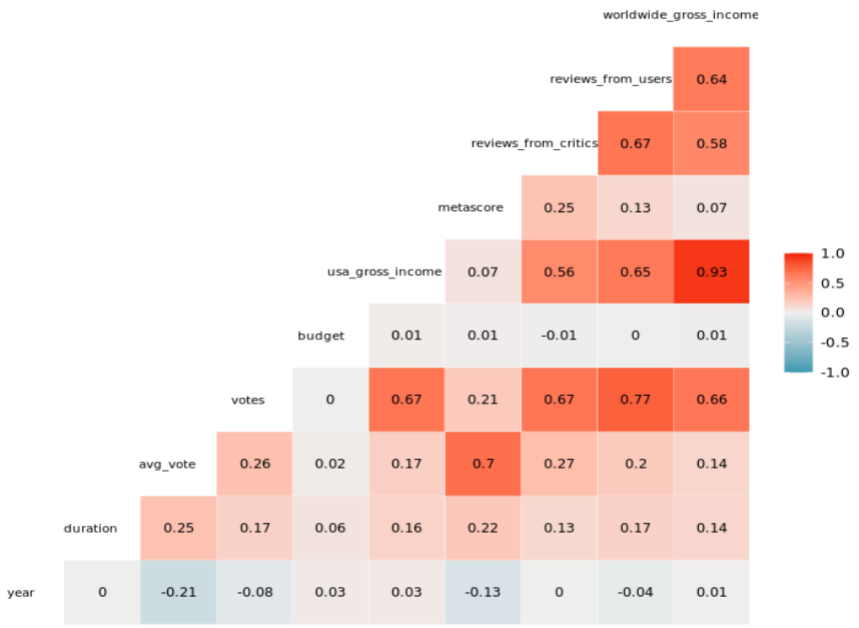
4.3 Статистические тесты

Что касается статистических тестов, для их анализа нам необходимы нулевая и альтернативная гипотезы, которые мы будем проверять. В нашем случае нулевая гипотеза: кассовые сборы не связаны с некоторой переменной, а альтернативная гипотеза: кассовые сборы связаны с этой переменной. При проведении данного теста в итоге мы получаем некое значение p-value. Если p-value большое (> 0.05), нулевая гипотеза принимается. Значит, нет оснований говорить, что между признаками вообще есть корреляция. Если маленькое, то принимается альтернативная.

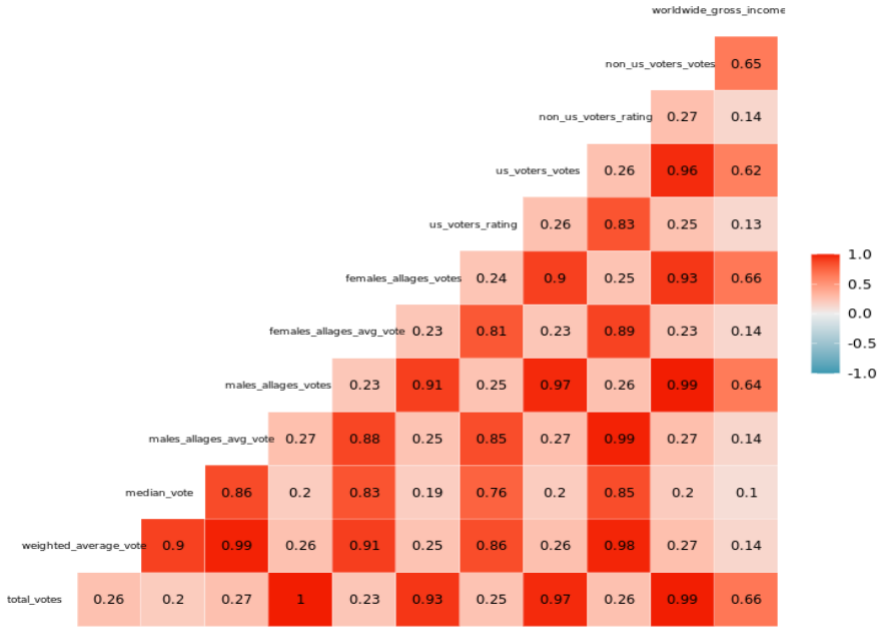
Для проведения статистических тестов мы привели переменные genre, language и country к длинному формату, то есть для каждого слова в тексте создали отдельную строку в данных. Это позволило нам сильно сократить количество уникальных наблюдений. Таким образом, мы провели 4 теста перестановок (зависимость кассовых сборов от жанра, страны, валюты и языка фильма), и все они показали маленькое значение p-value ($< 2.2e\text{-}16$). Значит, нулевая гипотеза отвергается. Эти тесты позволяют нам сделать вывод о наличии корреляции между признаками.

4.4 Коэффициенты корреляции

Для дальнейшего анализа мы разделили количественные переменные на 2 датасета, чтобы не перегружать графики корреляции. Мы рассчитали коэффициенты корреляции по Пирсону для всех числовых переменных, а потом для них же по Спирмену, так как распределение некоторых данных далеко от нормального и из-за этого коэффициент по Пирсону может давать неточные результаты, в то время как корреляция по Спирмену может гораздо точнее описать ситуацию.

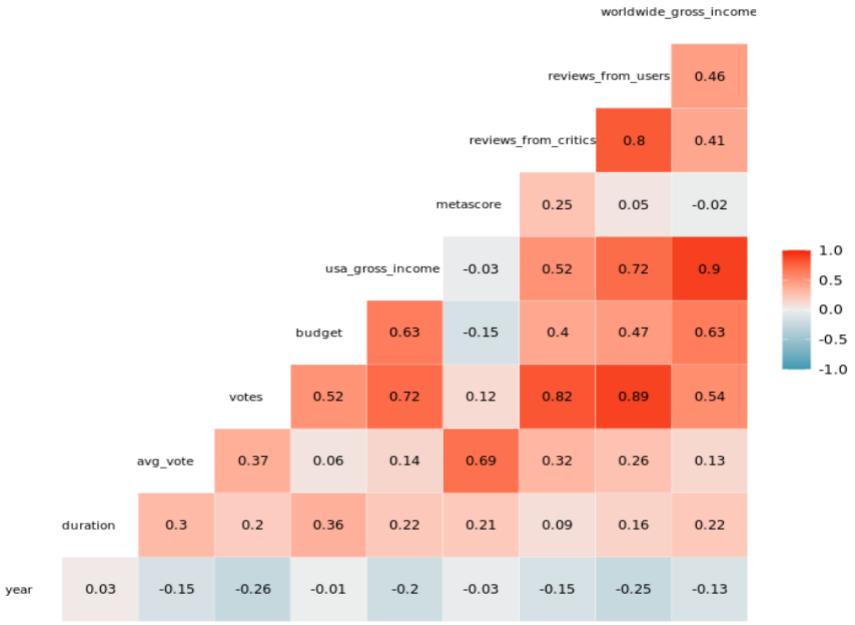


Коэффициент корреляции по Пирсону для первой части переменных

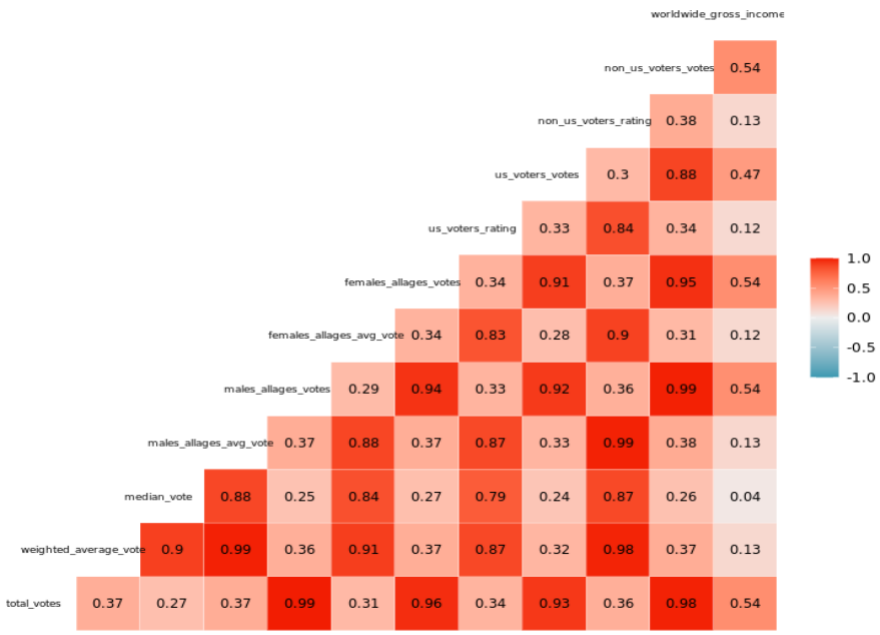


Коэффициент корреляции по Пирсону для второй части переменных

Таким образом, наибольшим коэффициентом корреляции по Пирсону для всемирных кассовых сборов обладает переменная usa_gross_income - кассовые сборы в США. Это легко объяснить: кассовые сборы киносетей в пределах США занимают наибольшую долю общих сборов из-за высокой стоимости билетов (для сравнения, средняя цена билета в США составляет чуть более $9, а в России - чуть больше $3.5) и количества кинотеатров (по данным ExtraCharts.com США занимает лидирующую позицию по количеству киноэкранов). Далее корреляцию более 65% для всемирных кассовых сборов имеют переменные votes (*количество голосов от пользователей*), total_votes (*общее количество голосов*), females_allages_votes (*среднее количество голосов от женщин*) и non_us_voters_votes (*количество голосов от пользователей не из США*), то есть почти все, что связанно именно голосами от пользователей. И это тоже несложно объяснить: вполне закономерно люди захотят смотреть фильмы с высоким рейтингом, полученным по большому числу голосов пользователей.



Коэффициент корреляции по Спирмену для первой части переменных

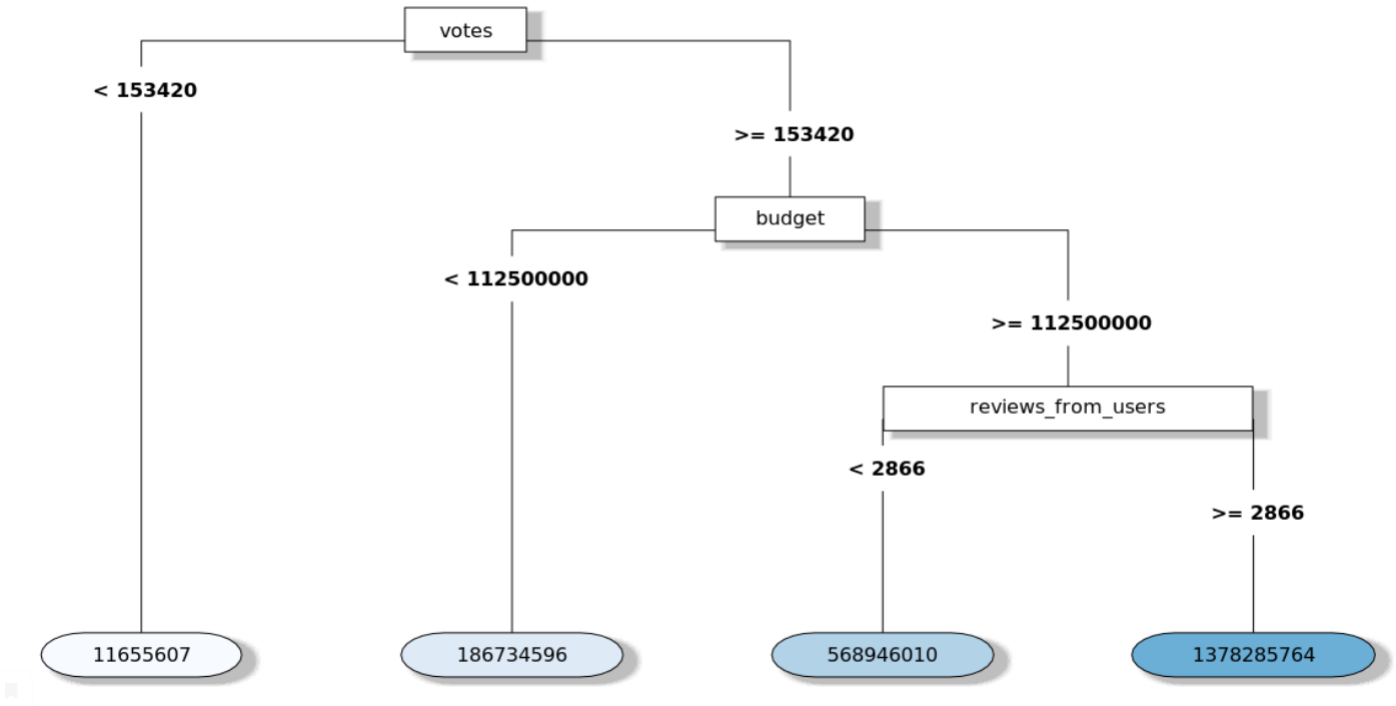


Коэффициент корреляции по Спирмену для второй части переменных

Что касается коэффициента корреляции по Спирмену (в отличие от коэффициента Пирсона, этот вариант коэффициента корреляции работает не с исходными значениями переменных, а с их рангами), наибольшее значение имеет также переменная usa_gross_income - кассовые сборы в США (даже при большей точности коэффициента корреляции по Спирмену мы получаем те же результаты: кассовые сборы в США играют решающую роль в общих сборах), а показатель немногим более 50% имеют переменные votes (*количество голосов от пользователей*), budget (*бюджет фильма*), total_votes (*общее количество отзывов*), females_allages_votes (*среднее количество голосов от женщин*), males_allages_votes (*среднее количество голосов от мужчин*) и non_us_voters_votes (*количество голосов от пользователей не из США*), то есть также зависимость от голосов в большей степени, но уже на порядок ниже, хотя при этом и появились новые переменные: budget **(0.63)**, reviews_from_critics (*количество голосов от критиков*) **(0.41)**. Действительно, бюджет фильма зачастую играет важную роль в общих сборах фильма. Производственный бюджет рационально расходуется на съемочный процесс, пост-производство и распространение (которое включает в том числе рекламу).

4.5 Предиктивная модель

Заключительной частью нашего анализа является предиктивная модель в виде дерева. Чтобы результаты были воспроизводимыми, используется функция set.seed(), которая каждый раз при использовании генератора псевдослучайных значений берет идентичные последовательности чисел. Таким образом, мы получаем небольшое дерево, в котором величина кассовых сборов по всему миру зависит в первую очередь от количества голосов пользователей, а потом уже от бюджета фильма и количества отзывов пользователей.



Предиктивная модель в виде дерева

Мы можем заметить, что не все физически значимые переменные из рассчитанных коэффициентов корреляции вошли в это дерево. Это объясняется тем, что дерево строится на неполной выборке, которая берется рандомно из начального датасета, тем самым деревья могут получаться разные, а переменные с высоким уровнем значимости могут заменять друг друга. Так что это лишь один из вариантов модели, построенный на выбранной нами выборке. Мы использовали set.seed(), чтобы исследование имело воспроизводимые результаты. Подобный метод анализа данных имеет множество преимуществ: полученные результаты легко интерпретируемы, требует малой подготовки данных и легко работает с большими набора данных (что особенно актуально в нашем исследовании). Следует понимать, что использованное нами значение функции дает наиболее релевантную и значимую интерпретацию данных.

5 Результаты

В данной главе мы рассказываем, к чему пришли в ходе проведения статистического анализа и оцениваем, как это можно использовать на практике или в будущих исследованиях.

Проведенный нами анализ доказывает поставленную гипотезу о существовании факторов, влияющих на кассовые сборы фильмов и определяемых с помощью предиктивных моделей и механизмов программирования. Использовав разные методы исследования, мы нашли некоторые закономерности, которые можно использовать в дальнейшем для более масштабных исследований на данную тематику. Например, с бóльшим количеством переменных. Не менее интересным выглядит исследование корреляции актерского состава и общих сборов картины. Наша работа как раз поможет облегчить будущие исследования на эту тематику.

Мы смогли показать и корреляцию кассовых сборов в зависимости от дня недели, когда выходит фильм: действительно, как показывают и наши тесты, и практика наиболее кассовые и успешные фильмы выходят в период со среды по пятницу. Не менее интересным стали и результаты исследования зависимости кассовых сборов от месяца выхода фильма. Зимний период (особенно декабрь и январь) оказались наиболее успешными для выпуска фильмов. И правда, большое количество праздничных дней, новые фильмы с новогодней и рождественской тематикой играют серьезную роль в создании бокс-офиса картины. Не стоит забывать и о большом количестве блокбастеров, которые как раз выходят в праздничные периоды. Кроме этого, мы смогли наглядно показать важнейшние периоды в развитии кинематографа: 30-40-е годы, связанные с появлением цветного кино, новых технологии съемки и культовых актеров и актрис; 80-90-е годы и связанные рост числа кинотеатров, подъем культуры кинематогрофа - кино становится одним из важнейших культурных показателей многих стран (Голливуд становится именем нарицательным); 2000-е, когда индустрия кино становится главным развлечением по всему миру.

Не менее важным результатом нашего исследования является изучение корреляции между отзывами на фильмы от женщин и мужчин и общими сборами картин. Это еще раз показывает важность так называемых тестовых показов - предрелизных показов фильмов, в ходе которых у выбранных зрителей собирают отзывы о качестве картины, ее недостатках и преимуществах. Наше исследование поможет киностудиям заранее формировать списки зрителей для получения наиболее релевантой оценки «усредненного» потребителя фильма. Это, в свою очередь, поможет избежать экономических потерь при прокате новых лент. Кроме того, наше исследование показывает, насколько важно для киностудий и дистрибьюторов агитировать зрителей оставлять отзывы на просмотренные фильмы.

6 Выводы

В данной главе мы делаем выводы из нашего исследования.

В ходе нашего исследования были выявлены наиболее существенные факторы, влияющие на кассовые сборы фильмов. Мы создали основу для более глобальных исследований. Нам удалось создать предиктивную модель, которая удачным образом может быть применена на практике, в том числе различными киностудиями и кинодистрибьюторами. Что более важно, нашим исследованием мы смогли доказать, что кинематограф - это комплексная и крайне точная индустрия, в которой важны все процессы: от создания идеи фильма до его проката. Использование предикативных моделей поможет облегчить эти процессы и исключить потенциально не прибыльные картины из проката или уменьшить их долю.