### Федеральное государственное автономное образовательное учреждение высшего профессионального образования

### Национальный исследовательский университет

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Факультет экономики

Магистерская программа

"Математические методы анализа экономики"

Кафедра математической экономики и эконометрики

##### МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

**Эконометрическое прогнозирование**

**кассового успеха кинофильмов**

Выполнил

Студент группы № 71-ММАЭ2 Педяш Евгений Анатольевич

Научный руководитель Доцент кафедры математической экономики и эконометрики,

К.ф.-м.н. Демидова Ольга Анатольевна

Москва 2013

Оглавление

Введение 3

Глава 1. Источники данных 6

Глава 2. Анализ кассовых сборов фильмов 8

2.1 Краткая методическая справка 9

2.2 Динамическая модель кассовых сборов фильмов 11

2.3 Детальное описание компонент динамической модели 12

2.4 Модели динамики объясняющих переменных 26

2.5 Модели индивидуальных отличий между фильмами 28

Глава 3. Прогноз кассовых сборов фильмов 32

3.1 Первый этап прогноза кассовых сборов фильмов 33

3.2 Модель общих кассовых сборов фильмов 35

3.3 Второй этап прогноза кассовых сборов фильмов 38

Заключение 39

Список использованной литературы, источники данных 40

Приложение 41

**Введение**

Кинематограф является неотъемлемой частью современной культуры. Со времен публичных демонстраций первых короткометражных фильмов в конце 1885г процесс создания кинокартин и их показ зрителям эволюционировали в отдельную отрасль экономики – киноиндустрию.

В данной работе исследовался классический канал распространения кинокартин: прокат фильмов в кинотеатрах. Целью данной работы было исследовать с помощью современных эконометрических подходов кассовые сборы фильмов и предложить подход для опережающего прогнозирования кассовых сборов новых фильмов.

История показала, что производство кинофильмов крайне рискованное занятие. На практике киностудии получают баснословные прибыли от проката нескольких «блокбастеров» и теряют миллионы долларов на производстве десятков «провальных» фильмов в год.

Стоит заметить, что успех фильмов в кинотеатрах напрямую связан с последующим успехом данных фильмов в других каналах распространения: на телевизионных каналах, на DVD и других носителях. Таким образом, вопрос прогнозирования кассовых сборов фильмов оказывается интересен не только киностудиям, дистрибьюторам фильмов и кинотеатрам, но и инвесторам киноиндустрии, представителям розничной торговли и представителям телевизионных каналов.

Правила рынка кинопроката достаточно просты и проблем с прогнозированием предложения фильмов обычно не возникает. Основная сложность в прогнозировании кассовых сборов фильмов возникает на стороне прогнозирования спроса на фильмы.

Множество ученых пробовали свои силы в статистическом прогнозировании кассовых сборов. Задача прогнозирования спроса на фильмы относится к классу задач большой размерности, поскольку список различных характеристик фильма значительно превосходит количество фильмов участвующих в анализе. Задачи большой размерности обычно требуют значительных упрощений в процессе решения, и общего мнения относительно данных упрощений среди ученых нет. В результате, разные ученые акцентируют свое внимание на разных аспектах прогнозирования спроса на фильмы и получают хоть и не противоречивые, но все же различные результаты.

В своей работе Ramesh S. и Dursun D. (2006) [1] обучили нейронную сеть предсказывать кассовые сборы фильмов. В работе использованы данные о жанре фильма, возрастных ограничениях фильма, уровне конкуренции в период проката, уровне звезд в фильме, уровне спецэффектов в фильме, количестве экранов на которых данный фильм был доступен для зрителей и данные о том является ли фильм продолжением истории другого фильма.

Marton M., Taha Y. и Janos K. (2012) [2] показали, что уровень активности пользователей на интернет страницах посвященных фильму может быть использован, в качестве прокси метрики «ажиотажа», для раннего предсказания кассового успеха фильма.

Взаимосвязь между кассовыми сборами фильмов и обзорами критиков относительно данных фильмов проанализировал Kennedy A. (2008) [3].

Особый интерес для киностудий может представлять работа Eliashberg J., Hui S. K. и Zhang Z. J. (2007) [4]. В данной работе с помощью статистических методов обработки текстовой информации выявлена взаимосвязь между содержанием сценария и кассовыми сборами фильма.

Наши соотечественники Antipov E. и Pokryshevskaya E. (2010) [5] пришли к выводу, что для корректного анализа кассовых сборов первой недели фильмов необходимо раздельно анализировать различные типы фильмов.

Работы Achal A., Manas P. (2011) [6] и Andrei O., Mathias B., Manos T., Maarten R. (2012) [7] несколько отличаются от перечисленных ранее работ. В них нет анализа кассовых сборов фильмов, но идейно они близки данной цели. В работах анализируется IMDb рейтинг фильмов – зрительская бальная оценка фильмов. Данная бальная оценка является зрительской оценки «качества» фильма и, следовательно, может использоваться при оценке спроса на фильм.

В своей работе Achal A. и Manas P. (2011) [6] создали алгоритм для прогноза рейтинга фильма на основе данных об участниках процесса производства кинокартины: актерах, режиссерах и сценаристах.

В работе Andrei O., Mathias B., Manos T. и Maarten R. (2012) [7] показана взаимосвязь между рейтингом фильма и реакцией на сообщения относительно данного фильма пользователей социальных медиа. Данная работа в чем-то схожа с работой Marton M., Taha Y. и Janos K. (2012) [2], в них обеих в качестве объясняющих переменных используются информационные метрики мониторинга интернета.

Спектр статистических методов, используемых в данных работах, может удивить даже опытного ученого: от простейших линейных регрессионных моделей до сложных иерархических моделей, от простейших логистических регрессионных моделей до сложных подходов с использованием нейронных сетей.

Во всем калейдоскопе подходов и методов прослеживаются две закономерности:

* подавляющее большинство работ по анализу кассовых сборов проведено на основе данных о кассовых сборах США. Даже наши соотечественники предпочитают анализировать данные США.
* во всех работах игнорируется динамический аспект кассовых сборов фильмов: либо анализируются кассовые сборы первой недели, либо анализируется общая сумма кассовых сборов фильмов.

В данной работе в качестве площадки для анализа были взяты данные по кассовым сборам фильмов в России за период с 2008 по 2012.

Преследовалась основная практическая цель – отобрать и адаптировать под реалии Российского кинопроката идеи, использованные в анализе кассовых сборов США и предложить подход для опережающего прогнозирования кассовых сборов фильмов в России.

Также преследовалась и другая, не менее важная, цель: проанализировать динамические аспекты кассовых сборов фильмов.

По результатам анализа динамики кассовых сборов фильмов найдено значимое влияние таких динамических аспектов как: конкуренция между фильмами в прокате, влияние праздников и реклама дистрибьюторов фильмов. Также найдено, что скорости падения кассовых сборов фильмов различны, различия значимы и частично объясняются характеристиками фильмов.

Практическая новизна данной работы заключается в разработке подхода для опережающего прогнозирования кассовых сборов фильмов для России. Научная новизна данной работы заключается в произведенном анализе динамики кассовых сборов.

**Глава 1. Источники данных**

Динамические характеристики (из открытых источников):

1. **Объем кассовых сборов (в динамике по неделям)** – данные взяты из Бюллетеня Кинопрокатчика [10], специализированного сайта для представителей Российской киноиндустрии
2. **Цена билетов (в динамике по неделям)**, средняя цена билета за уикенд **-** данные взяты из Бюллетеня Кинопрокатчика [10]
3. **Количество копий (в динамике по неделям)**, количество экранов, на которых демонстрировался фильм **-** данные взяты из Бюллетеня Кинопрокатчика [10]
4. **Праздники (в динамике по дням)** - данные взяты из Производственного календаря РФ

Динамические характеристики (из закрытых источников информации):

1. **Объем рекламы (в динамике по неделям**) - данные взяты из мониторинга рекламной активности TNS Gallup: в качестве объема рекламы на телевидении использовались недельные рейтинги на самую общую аудиторию, в качестве объема рекламы в других медиа использовались оценки бюджетов потраченных на данные медиа. Оценка отложенного влияния рекламы проводилась по закрытой методике компании BIG. Поскольку вопрос оценки эффективности различных медиа не является центральным в данной работе, то в дальнейшем рассказе под «объемом рекламы» будем понимать единый фактор сформированный из всех медиа затрат, методика построения которого дана нам извне.
2. **Влияния праздников на кассовые сборы (в динамике по неделям)** -данные получены при анализе внутри-недельной динамики объема кассовых сборов. Объем кассовых сборов по дням взят из закрытого источника Rentrak (Астра Пэйдж). Преимущество данного источника информации: высокая детализация динамики. Основной недостаток: данный мониторинг доступен не по всем фильмам. Как производилась оценка влияния праздников на кассовые сборы недели, будет упомянуто в основной части описания модели. Как и в случае рекламы, под «влиянием праздников на кассовые сборы недели» можем понимать единый фактор, сформированный из данных производственного календаря РФ, методика построения которого дана нам извне.

Индивидуальные характеристики фильмов (из открытых источников):

1. **IMDb рейтинг фильма (байесовская оценка среднего, взвешенная к аудитории кинотеатров США)**, пользовательская оценка «качества» фильма по 10бальной шкале – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
2. **Жанр фильма**, фильм может быть классифицирован одновременно к нескольким жанрам – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
3. **Страна производства фильма**,ISO коды всех стран участвовавших в производстве картины **-** данные взяты из Бюллетеня Кинопрокатчика [10]
4. **Взаимосвязи с другими фильмами**, является ли фильм продолжением, переделкой или пародией на другой фильм, содержит ли ссылки на другие фильмы – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
5. **Длительность фильма (или длинна 35мм носителя)** – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
6. **Технические параметры фильма**, является ли фильм широкоформатным, на каком оборудовании производилась сьемка – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
7. **Кинолаборатории задействованные в производстве фильма** – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
8. **Дистрибьютор фильма в России -** данные взяты из Бюллетеня Кинопрокатчика [10]
9. **Дата премьеры фильма в России -** данные взяты из Бюллетеня Кинопрокатчика [10]
10. **Дата премьеры фильма в США** – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
11. **Форматы релиза фильма Российском прокате**, 35мм/цифровой/3D **-** данные взяты из Бюллетеня Кинопрокатчика [10]
12. **Форматы релиза оригинала фильма**, 35мм/цифровой/3D – данные взяты из базы данных The Internet Movie Database (IMDb) [9]
13. **Возрастные ограничения фильма**, 18+, вход с сопровождением родителей, без ограничений – данные взяты из базы данных The Internet Movie Database (IMDb) [9]

**Глава 2. Анализ кассовых сборов фильмов**

Данная глава будет посвящена анализу кассовых сборов фильмов.

В начале главы приводится краткая методическая справка об основном статистическом методе, использованном в данной работе при анализе кассовых сборов фильмов.

Далее сформулирована спецификация динамической модели кассовых сборов фильмов и приводится детальное описание всех компонент данной модели.

В заключение главы описаны две дополнительные надстройки к динамической модели кассовых сборов: модели динамики объясняющих переменных и модели индивидуальных отличий между фильмами – эти надстройки позволят использовать динамическую модель кассовых сборов для опережающего прогнозирования кассовых сборов новых фильмов.

**2.1 Краткая методическая справка**

**Модель со случайными коэффициентами (Mixed-eﬀects model)**

Модель со случайными коэффициентами относится к классу методов, используемых в анализе панельных данных. Данная модель позволяет одновременно включать параметры с фиксированными эффектами и параметры со случайными эффектами.

В общем виде модель со случайными коэффициентами можно записать:

,где и - фиксированные эффекты модели;

независимые одинаково распределенные ошибки;

и – случайные эффекты модели .

Обычно полагают, что случайная константа и случайный наклон подчиняются двумерному нормальному закону распределения. Связаны некоторой ковариационной матрицей и соответствующим коэффициентом корреляции между случайной константой и случайным наклоном .

Модели данного типа оцениваются методом максимального правдоподобия. Результаты оценки модели в значительной степени зависят от первичных предположений относительно ковариационной матрицы .

**2.2** **Динамическая модель кассовых сборов фильмов**

**Спецификация модели**

Почти каждая компонента динамической части модели потребует комментариев, поэтому для каждой компоненты модели будет отведён отдельный небольшой раздел. А для начала запишем динамическую модель с кратким описанием переменных:

,где ***i***– индекс фильма; ***t***– номер недели в прокате для данного фильма;

**-** индекс кассовых сборов фильма ***i*** на неделе ***t***;

**–** рейтинг качества фильма ***i***, зрительская оценка по 10-ти бальной шкале;

**–** копии,количество экранов на которых демонстрировался фильм ***i*** на неделе ***t***;

**-** индекс цены фильма ***i*** на неделе ***t***;

**–** влияния праздников на неделе ***t***;

**–** влияния рекламы фильма ***i*** на неделе ***t***;

**–** фактор присутствия фильмов-конкурентов ***j*** на неделе ***t***;

**–** различия между США и Россией в датах выхода фильма ***i*** в прокат;

**, , ,,,,,,, –** константа, тренд и остальные коэффициенты, отражающие фиксированное влияние соответствующих параметров на кассовые сборы всех фильмов;

и **–** ненаблюдаемые случайные компоненты константы и тренда кассовых сборов фильма ***i .***

**2.3 Детальное описание компонент динамической модели**

**Фильм (*i=1,…742)***

**Фильм** является основной единицей «товара» предоставляемого кинотеатрами, поэтому крайне важно понимать какие фильмы оказались предметом анализа в данной работе.

По общим данным Бюллетеня Кинопрокатчика [10] о фильмах, в период с 2008 по 2012 год в Российский кинопрокат вышли чуть более 1600 фильмов.

Детальные данные с динамикой кассовых сборов в Бюллетене Кинопрокатчика [10] можно найти только в форме недельных отчетов о двадцати наиболее успешных фильмах недели – двадцати фильмов с наибольшими кассовыми сборами. В данных отчетах, так или иначе, фигурировали 1200 фильмов из 1600, оставшиеся 400 фильмов были не столь успешны в прокате.

По естественным причинам, различные фильмы в статусе «двадцати наиболее успешных» находились разное количество недель. В результате, собранная информация о кассовых сборах 1200 фильмов носила характер сильно не сбалансированных панельных данных. Для корректного анализа данных такой природы пришлось ввести дополнительных фильтр: фильмы с менее чем тремя динамическими наблюдениями в анализ не включались. Таким образом, в анализ были отобраны только 742 фильма – фильма продержавшихся более двух недель в списке двадцати лучших по сборам.

Среднее значение суммы кассовых сборов фильмов в анализируемой выборке – 187,5 млн. руб. на фильм. Среднее значение суммы кассовых сборов фильмов для исключенных из анализа фильмов – 9,5 млн. руб. Средние значения отличаются почти в 20 раз, что указывает на сильное «смещение отбора» динамически анализируемой выборки по сравнению с генеральной совокупностью всех фильмов выходивших в период с 2008 по 2012 год.

Данное «смещение отбора» навязано нам структурой открытых данных о кассовых сборах фильмов в России. При построении прогноза кассовых сборов новых фильмов необходимо иметь в виду наличие данного смещения, другими словами: модель обученная на данных «успешных» фильмов не имеет представления о существовании «не успешных» и при построении прогноза такая модель будет склонна преувеличивать кассовые сборы «не успешных» фильмов.

**Недели в прокате (*t=0,…13)***

В индустрии кинопроката время принято считать «викэндами», т.е. периодами недели от дня недели премьер до воскресенья. В России премьеры фильмов происходят по четвергам. Соответственно «викендом» в Российском прокате принято считать период с четверга по воскресенье, в этот период собирается основная часть кассовых сборов недели.

Использование «викэндов» в статистике не корректно, так как сумма кассовых сборов по всем «викендам» не равна (строго меньше) общей суммы кассовых сборов. Поэтому в нашем анализе для сборов фильмов будут использоваться полные **недели в прокате**, то есть периоды с четверга календарной недели по среду следующей календарной недели.

**Недели в прокате** для каждого фильма будут нумероваться, начиная с недели премьеры фильма. В случае если у фильма были предпоказы до официальной премьеры фильма, то в анализ кассовых сборов эти недели не включались.

Средняя динамика кассовых сборов первых 4 недель проката представлена в Приложении [Диаграмме 1]. С каждой следующей неделей кассовые сборы фильма теряют почти половину сборов текущей недели. Основная часть кассовых сборов фильмов в Российском кинопрокате приходится на первые 3 недели проката.

При такой короткой динамике сборов, кажется естественным упрощение: анализировать только отличия между фильмами – видимо такой логикой придерживались авторы перечисленных во введении статей [1-7].

Приглядевшись чуть более внимательно к [Диаграмме 1], можно увидеть некоторые различия в динамике кассовых сборов фильмов различных дистрибьюторов. Сборы фильмов от дистрибьютора Central Partnership убывают медленней среднего, сборы фильмов от дистрибьютора Sony убывают быстрее среднего.

Если мы вернемся к модели, то процесс убывания кассовых сборов всех фильмов представлен фиксированным негативным трендом . Индивидуальные отличия фильмов по скорости падения кассовых сборов представлены случайной компонентой тренда .

**Индекс кассовых сборов фильмов ()**

Поскольку в анализе кассовых сборов участвовали фильмы за достаточно долгий период: с 2008 по 2012год, то возникают естественные сомнения о сопоставимости сборов разных лет.

Для приведения кассовых сборов разных периодов к сопоставимому базису, был рассчитан показатель инфляции для рынка кинопроката. Для этого была исследована макро-динамика средней цены билета и средней взвешенной на количество копий цены билета, иллюстрация макро-динамики цен приведена в Приложении на [Диаграмме 2]. Из предположения о постоянном темпе изменения цен была рассчитана оценка годового показателя инфляции для рынка кинопроката, которая составила 10.3%, и была построена средняя инфлированная цена билета.

Как видно на [Диаграмме 2] полученная средняя инфлированная цена билета неплохо описывает макро-динамику и средних и средних взвешенных на количество копий цен билетов.

Индекс кассовых сборов отдельных фильмов формировался делением кассовых сборов (в номинальном выражении) на среднюю инфлированную цену билета, таким образом, кассовые сборы фильмов были приведены к сопоставимому базису. Аналогичным образом к сопоставимому базису были приведены цены билетов отдельных фильмов.

В качестве зависимой переменной модели использовалась логистическая трансформация индекса кассовых сборов. Данная трансформация используется для борьбы с «тяжелыми хвостами» в распределении индекса кассовых сборов фильмов и для приведения модели к мультипликативной форме. Идея логистической трансформации кассовых сборов в той или иной мере участвует в работах [1],[2] и [5], поэтому нет смысла отдельно останавливаться на данном вопросе.

**Imdb рейтинг фильмов (**

Как уже несколько раз упоминалось ранее, Imdb рейтинг это средняя взвешенная зрительская оценка качества фильма (в 10 бальной шкале). Безусловно, Imdb рейтинг не является уникальным, почти у каждого локального ресурса о фильмах есть свой собственный зрительский рейтинг фильмов – поэтому выбор именно данного рейтинга в качестве прокси «качества» фильма требует обоснования.

В рейтинге Imdb исследуемые в данной работе фильмы были представлены наиболее полно.

Imdb рейтинг - самый цитируемый рейтинг фильмов в мире. Немалое количество научных работ написано на тему его предсказания. Что в перспективе дает возможность прямого использования данных подходов для его предсказания.

Данный рейтинг сформирован в виде байесовской оценки среднего и по заявлениям авторов устойчив к таким внешним попыткам изменения оценки как, например, заказным голосованиям со стороны пиар агентств.

При формировании Imdb рейтинга зрительские оценки «качества» фильма приводятся (взвешиваются) к аудитории зрителей США, что может слабо согласоваться с мнением Российских зрителей о «качестве» данного фильма. С другой стороны, детальный анализ дат выхода фильмов в прокат показывает, что более четверти анализируемых фильмов в Российский прокат попадают с более чем месячным опозданием после премьеры в США – для этой группы фильмов Imdb рейтинг может использоваться в качестве опережающего индикатора «качества» фильмов.

**Копии (**

Основной мерой предложения на рынке кинопроката являются **копии**. Под **копиями** понимаются суммарное количество копий фильмов в виде физических комплектов 35мм пленки и копий фильмов на цифровых носителях, выданных кинотеатрам от дистрибьюторов. Другими словами это максимальное количество экранов, на которых данный фильм мог быть легально показан единовременно всеми кинотеатрами.

Копии распространяются посредством двухсторонних договоренностей между дистрибьюторами и кинотеатрами. В общем виде данные договоренности сводятся к тому, что за символическую плату кинотеатр получает копии (одну или несколько) с правом показа данного фильма в течение следующих 2 недель и по итогам этих двух недель обязуется значительную долю кассовых сборов вернуть дистрибьютору. У кинотеатра есть возможность продлевать данную договоренность и возможность докупать дополнительные копии.

К сожалению, статистика по копиям агрегирована и не содержит данных о структуре копий (нет данных о долях пленочных копий, цифровых и 3D копий фильма); не содержит данных о количестве реальных показов фильма кинотеатрами.

Форма зависимости логарифма индекса кассовых сборов от логарифма количества копий носит выраженный не линейный характер (представлена в Приложении на [Диаграмме 3]).

Природа такой зависимости, скорее всего, связана с неоднородностью в структуре копий формируемых большими сетями кинотеатров и кинотеатрами одиночками.

Возможности проверить данную интерпретацию у нас нет поскольку нет детальных данных о составе копий. И интерпретация остается сформулированной в виде гипотезы.

**Индекс цен фильмов (**

Цены билетов фильмов были поделены на среднюю инфлированную цену билета. Таким образом, был сформирован индекс цен фильмов, сопоставимый между всеми анализируемыми фильмами.

Изменения цен фильмов могут быть вызваны множеством различных эффектов. Цены между фильмами могут отличаться из-за разницы в форматах между этими фильмами - классическое 35мм кино, цифровое кино или 3D. Цена фильма может меняться в динамике из-за перемещения показов фильма внутри дня - кинотеатры могут перемещать фильм из дорогих вечерних сеансов в более дешевые дневные сеансы. Цены также могут отличаться из-за региональной структуры кинотеатров закупивших копии фильмов – некоторые фильмы могли показываться исключительно в дорогих столичных кинотеатрах (закрытые показы).

Разделить влияние всех перечисленных эффектов на цену не представляется возможным, так как цена, предоставляемая в открытых источниках, агрегирована до уровня средней цены билета по России - перечисленные эффекты уже смешаны.

**Праздники (**

Праздники являются одной из основных причин формирования сезонности во многих процессах современной жизни людей.

В праздники экономическая жизнь многих отраслей экономики практически останавливается, но индустрия развлечений – и киноиндустрия в частности – являются исключениями из данного правила.

Общие кассовые сборы киноиндустрии могут практически удваиваться в праздничные дни. Кинокомпании специально учитывают праздники при планировании дат премьер своих лучших фильмов, дистрибьюторы учитывают праздники при планировании рекламной поддержки собственных фильмов, кинотеатры готовятся к наплыву зрителей в праздничные дни – все это приводит к значительному росту конкуренции между фильмами в период праздников.

Для корректного учета влияния праздников была проанализирована динамика кассовых сборов по дням. Для этого были взяты данные о кассовых сборах по дням из закрытого источника Rentrak (Астра Пэйдж). На основе этих данных была оценена динамическая модель кассовых сборов. Модель была полностью эквивалентная указанной выше, за исключением некоторых отличий: а) части параметров – доступных только по неделям – в модель не включалась; б) индекс ***t*** в этой модели представлял дни, а не недели; в) для корректной оценки внутри-недельной сезонности кассовых сборов, была добавлена группа дамми переменных для различных дней недели; г) была добавлена дамми переменная для праздничных дней.

В модели отдельно тестировалось влияние «не полных рабочих дней» и «рабочих суббот и воскресенья», но значимого отличия в кассовых сборах данных дней обнаружено не было.

Результаты оценки внутри-недельной сезонности и оценки влияния праздников приведены в Приложении на [Диаграмме 4]. Из диаграммы видно что кассовые сборы «викенда» (с четверга по воскресенье) составляют значительную часть (74%) всех кассовых сборов типичной недели без праздников. Это объясняет, почему работники киноиндустрии мыслят в терминах кассовых сборов «викендов».

Из диаграммы также видно, что кассовые сборы праздничного дня более чем в два раза превосходят сборы аналогичного дня без праздника. Отдельно стоит отметить, что поскольку график праздничных дней был взят из Трудового календаря РФ, то по логике данного источника – праздничный день не может попадать на субботу и воскресенье, в таком случае он переносится на другой рабочий день.

Для учета влияния праздников, в недельной модели кассовых сборов фильмов, был сформирован фактор. По сути, данных фактор является арифметической суммой надбавок (из [Диаграммы 4]) за все праздники недели.

Например:

* Если в неделе проката нет праздников – фактору присваивалось значение **=0**
* Если в неделе проката в пятницу был праздник – фактору присваивалось значение **=0.17**
* Если в неделе проката все дни были праздничные (новогодние выходные) – фактору присваивалось значение

**=0.13+0.17+0.11+0.11+0.10=0.62**

**Реклама фильма (**

Данные об объеме рекламы были взяты из коммерчески распространяемого источника – мониторинга рекламы TNS Gallup. Эти данные были обработаны по закрытой методике компании BIG, так был сформирован единый фактор отражающий влияние рекламы на кассовые сборы. В данной работе будут описаны смысл и логическое обоснование необходимости данного преобразования, но технические подробности методики не будут раскрыты.

Процесс мониторинга рекламной активности состоит их двух основных этапов: 1) на первом этапе фиксируются факты выхода рекламных сообщений в основных каналах коммуникации; 2) по ценникам (взятым у представителей данных каналов) строится оценка рекламного бюджета потраченного на данные рекламные сообщения.

Цены у разных каналов коммуникации напрямую связаны с размером аудитории (количеством зрителей) потребляющей данный канал.

В результате, рекламный бюджет, оцененный по описанной выше методике, может быть использован как прокси мера размеров аудитории видевшей рекламное сообщение.

В киноиндустрии значительная часть рекламных сообщений размещается до премьеры фильма, так дистрибьюторы фильмов пытаются заранее привлечь зрителей.

С одной стороны, в модели мы можем использовать не модифицированные рекламные бюджеты в разбивке по неделям. Но таким образом мы не учтем эффект рекламных сообщений произошедших до премьеры фильма, поскольку их бюджеты не попадают в анализируемый период.

С другой стороны, мы можем использовать в модели сумму всех рекламных бюджетов, потраченных к данной неделе. Но таким образом мы в неявной форме делаем предположение о том, что реклама, сделанная за год до премьеры, эквивалентна рекламе, сделанной в неделю премьеры.

Обе альтернативы не реалистичны. В первой альтернативе сделано предположение о полном отсутствии памяти в аудитории рекламных сообщений, во второй альтернативе сделано предположение об идеальной памяти в аудитории рекламных сообщений – в реальной жизни мы имеем дело с чем-то средним.

Единый фактор, отражающий влияние рекламы на кассовые сборы, сформирован из тех соображений, что аудитория рекламных сообщений обладает памятью, но эта память не идеальна.

**Конкуренция**

Придя в кинотеатр, зритель сталкивается с проблемой выбора. Не многие зрители готовы за одно посещение кинотеатра просмотреть подряд все доступные фильмы. Поэтому вопрос о внутренней конкуренции между фильмами – одновременно находящимися в прокате - стоит особенно остро.

Количество копий фильмов конкурентов () может использоваться в качестве меры количества предложения конкурентов.

Средний рейтинг может использоваться в качестве меры среднего качества предложения конкурентов.

Обе предложенные меры конкуренции неплохо показали себя в процессе подбора модели описывающей кассовые сборы фильма. Если объединить в один фактор эти два показателя, то модель кассовых сборов значимо улучшается – улучшается в смысле значения функции правдоподобия и в смысле значений информационных критериев Шварца и Акаике (BIC, AIC).

В результате, в качестве меры общей конкуренции использовался сложный показатель, сформированный как сумма из отдельных влияний фильмов конкурентов. В качестве отдельного влияния фильма конкурента использовалось произведение количества копий фильма на рейтинг фильма. Сумма производилась по десяти лучшим (в смысле кассовых сборов) фильмам недели.

Идея данного показателя достаточно проста:

* если фильм конкурент показывается на малом количестве экранов, то каким бы качественным он не был – мы считаем, что этот фильм создает слабую конкуренцию;
* если фильм конкурент низкого качества, то на каком бы количестве экранов он не был показан – мы считаем, что этот фильм создает слабую конкуренцию;
* если фильм конкурент показывается на большом количестве экранов и при этом является качественным – мы считаем, что этот фильм создает высокую конкуренцию.

По логическим соображениям, в полученный латентный показатель конкуренции не должен включаться непосредственно моделируемый фильм, те есть .Так мы избежим проблем с эндогенностью в модели. Именно такой показатель и использовался во всех моделях.

С другой стороны, если мы посчитаем данный показатель по всем фильмам из десяти лучших (без исключений), то мы получим динамический показатель, не привязанный к какому-либо конкретному фильму. Данный показатель мы можем считать латентным показателем общей конкуренции между фильмами данной недели – динамика показателя в Приложении на [Диаграмме 5]. Из данной диаграммы видно, что конкуренция между фильмами растет на протяжении исследованного периода времени. Рост конкуренции обусловлен увеличением количества копий фильмов, при неизменном уровне качества фильмов.

**Разница в датах релиза фильма в России и США**

Разница в датах релиза фильма в России и США была разделена на два взаимоисключающих случая: запаздывание релиза в России по сравнению с США и опережение релиза в России по сравнению с США. Поскольку причины возникновения этих ситуаций различны, то и эффекты описываемые соответствующими параметрами могут отличаться.

Если Российский релиз фильма запаздывает по сравнению с релизом США то падает спрос на данный фильм. Связано это с тем что за время запаздывания часть спроса будет удовлетворена за счет альтернативных источников распространения фильмов, чем больше запаздывание – тем больше потери в кинозрителях. Примерами альтернативных источников распространения фильмов могут служить: DVD копии фильма, показы на телевидении, цифровые копии (как легальные так и не легальные). Даже информационные сайты о фильмах могут в некоторой мере являться источником распространения фильмов - на таких информационных ресурсах зачастую можно столкнуться с информацией раскрывающей часть сюжетной линии фильма, что безусловно снижает желание смотреть данный.

Для корректного учета влияния альтернативных источников распространения фильмов надо использовать метрики объема потребления фильмов в данных источниках распространения, например такие метрики: количество проданных DVD, количество скачиваний фильма в интернете. Таких данных либо нет в открытом доступе либо они доступны в крайне не структурированном виде. Поэтому запаздывание релиза кинофильмов в России по сравнению США использовалось в качестве грубого прокси данного влияния.

Если Российский релиз фильма опережает релиз США то это может быть связано с содержанием фильма. В данную группу фильмов в основном попали фильмы отечественного производства и фильмы европейского производства, то есть фильмы локального производства. Данные фильмы обычно ориентированы на локального зрителя и часто уступают Голливудским картинам в качестве съемки. К показу в США допускаются не все фильмы локального производства, но только наиболее адаптированные под массовую аудиторию. Чем более адаптирован фильм под массовую аудиторию, тем быстрее его переведут на английский язык и покажут в кинотеатрах США. Таким образом, опережение релиза в России по сравнению с релизом США может быть использовано в качестве прокси меры адапрированности фильма под широкую аудиторию.

**Случайные компоненты кассовых сборов ( и )**

Прежде чем обсуждать смысл случайных компонент константы итренда кассовых сборов, скажем пару слов о причинах выбора данной спецификации модели.

В нашем распоряжении есть два наиболее популярных семейства методов оценки вариационных компонент панельных моделей: методы дисперсионного анализа (ANOVA) и методы максимального правдоподобия (MLE) – речь идет не о различных панельных моделях (Pooled, FE, RE), речь идет о различных методах оценивания данных моделей. Например, RE модель можно оценивать двумя способами:

1) с помощью двух шаговой процедурой FGLS. На первом шаге строим оценку ковариационной матрицы ошибок, на втором шаге строим GLS оценку коэффициентов модели (используя ковариационную матрицу ошибок из первого шага). Полученные оценки коэффициентов модели можно записать, как некоторое матричное преобразование X и Y – это ANOVA оценки.

2) с помощью функции правдоподобия. Формулируем гипотезы об ошибках в виде функции правдоподобия. Решаем оптимизационную задачу (ищем максимум функции правдоподобия), так мы получаем оценки коэффициентов модели – это MLE оценки.

Модель одна, но методы оценки разные и результаты оценки тоже разные.

В случае сбалансированных панелей, оценки ANOVA являются лучшими квадратичными несмещенными оценками (BQU) вариационных компонент, поэтому они являются предпочтительными. В случае не сбалансированных панелей, оценки MLE дают лучшие оценки вариационных компонент (в нашем случае, и .

Поскольку в данной работе мы имеем дело с данными имеющими структуру не сбалансированной панели, то для моделирования кассовых сборов фильмов использовались исключительно методы оценки из семейства методов максимального правдоподобия.

Более подробное сравнение различных методов оценивания в условиях несбалансированных данных можно прочитать в лекциях Markus M [12].

В частности, для подтверждения применимости Mixed моделей в анализе несбалансированных данных приводится экономическая работа Erik B, Kjersti-Gro L и Terje S [11] о неоднородности технологий внутри производственных функций.

Таким образом, в данной работе для анализа несбалансированных данных о кассовых сборах фильмов использовалась модель со случайными коэффициентами.

В Приложении приведена [Таблица 1] со сводными статистиками различных спецификаций моделей со случайными коэффициентами. Для сопоставления приведена спецификация RE модели, оцененной методом максимального правдоподобия и из таблицы видно, что данная спецификация полностью эквивалентна Mixed модели со случайной константой .

Из [Таблицы 1] видно, что лучшей - в смысле значения функции правдоподобия – является спецификация модели со случайной константой , трендом и ненулевой корреляцией между ними. Данная спецификация модели была выбрана в качестве основной.

Результаты оценки случайных параметров модели приведены в Приложении на [Таблице 2]. Из таблицы видно, что между оценками случайной константы и оценками тренда фильма наблюдается слабая отрицательная связь, точечная оценка коэффициента корреляции = -0.24.

Найденная слабая зависимость имеет достаточно простую интерпретацию: фильмы, имеющие высокие (при прочих равных) кассовые сборы, имеют тенденцию к быстрому падению сборов от недели к неделе.

Такая зависимость может объясняться тем, что общая аудитория посетителей кинотеатров в стране ограничена. Такое ограничение слабо влияет на мало популярные фильмы, но может быть существенным для популярных фильмов.

**Фиксированные влияния параметров модели**

**( , , ,,,,,,,)**

Результаты оценки динамической модели приводятся в Приложении на [Таблице 3]. Из таблицы видно, что все параметры значимы и знаки при всех коэффициентов согласуются с нашими представлениями о формировании кассовых сборов фильмов:

* Недели в прокате. Коэффициент при переменной временного тренда строго отрицательный, .Это означает, что при прочих равных средний фильм с каждой следующей неделей теряет 37.5% кассовых сборов предыдущей недели.
* Imdb рейтинг фильмов (по 10 бальной шкале). За каждую дополнительную единицу зрительского рейтинга фильм получает 8.2% к кассовым сборам каждой недели
* Копии. Зависимость кассовых сборов от количества копий характеризуется положительной отдачей от масштаба. Увеличение количества копий недели на 5%

1. приводит к росту кассовых сборов недели на 5% - при изначальном низком количестве копий (250)
2. приводит к росту кассовых сборов недели на 5.7% - при изначальном среднем количестве копий (500)
3. приводит к росту кассовых сборов недели на 6.3% - при изначальном высоком количестве копий (1000)

* Индекс цен фильмов. Кассовые сборы фильма напрямую зависят от цены билета. Эластичность кассовых сборов по цене равна **0. 84**, то есть изменение цены на 1% приводит к увеличению кассовых сборов на 0.84%.
* Праздники значимо положительно влияют на кассовые сборы фильма.
* Реклама значимо положительно влияет на кассовые сборы фильма.
* Конкуренция. Построенный латентный показатель конкуренции значимо негативно влияет на кассовые сборы фильма. При среднем для 2012года уровне конкуренции, увеличение количества копий среднего по качеству конкурента (imdbrating = 6) на 100 копий приводит к снижению кассовых сборов на 0.8%.
* Разница в датах релиза фильма в России и США. Любые отличия в датах релиза фильмов между Россией и США негативно сказываются на кассовых сборах фильма. Фильмы, вышедшие в Российском прокате с годовым опережением, собирают на 40% меньше. Фильмы, вышедшие в Российском прокате с годовым опозданием, собирают на 35% меньше.

С точки зрения кассовых сборов фильма – предпочтительны мировые премьеры фильмов.

Полученная модель неплохо описывает динамику кассовых сборов фильмов. Научная новизна данной работы заключается в приведенном анализе динамики кассовых сборов фильмов. В оставшейся части работы речь пойдет о решении практической задачи – прогнозе кассовых сборов фильмов.

Описанная выше модель позволила оценить ненаблюдаемые случайные компоненты константы и тренда кассовых сборов фильмов, а также позволила получить оценку влияния пользовательского рейтинга на кассовые сборы фильма.

В общем случае, при построении прогноза кассовых сборов новых фильмов параметры не будут известны для новых фильмов. Позднее мы построим прогнозные модели индивидуальных отличий () между фильмами.

**2.4 Модели динамики объясняющих переменных**

Динамическая модель кассовых сборов неплохо описывает динамику кассовых сборов фильмов. Но она ещё не приспособлена для прогнозирования, значения части объясняющих переменных данной модели не доступны в условиях опережающего прогноза кассовых сборов новых фильмов – модель требует некоторых доработок.

Для начала опишем типичную структуру доступной информации о предстоящих премьерах фильмов. Потом, исходя из структуры данной информации, мы сформулируем реалистичную задачу для прогноза кассовых сборов данных фильмов.

Наиболее полную информацию о предстоящих премьерах мы можем найти в графике релизов. График релизов состоит из: наименований фильмов (по наименованию фильмов мы можем собрать данные о наблюдаемых индивидуальных характеристиках 7-18 этих фильмов); примерных дат релизов фильмов; дистрибьюторах фильмов и количестве копий первой недели релиза фильмов.

График релизов составлен примерно на 2 месяца наперед, постоянно обновляется. Но данные о копиях доступны с достаточной точностью только на 5-6 недель наперед, следовательно, реалистичным горизонтом прогнозирования будем считать прогноз на 6 недель.

При построении прогноза, у нас нет данных об объеме рекламной активности дистрибьюторов фильмов. Попытки прогноза динамики данного показателя не были успешны, поэтому в динамических прогнозах кассовых сборов фильмов предполагался средний уровень рекламного присутствия для каждого дистрибьютора.

При построении прогноза, отсутствуют данные о ценах фильмов. Индекс цены билета в равной степени может считаться как динамической, так и индивидуальной характеристикой фильма, within-вариация и between-вариация индекса цен примерно равны.

Попытки прогноза динамики индекса цен фильмов не принесли успеха. Лучший коэффициент детерминации, достигнутый для динамики индекса цены, не превысил значения . Поэтому в прогнозе мы будем предполагать, что цены фильмов статичны и отличаются только между фильмами. Модель для прогноза индекса средних цен фильмов приведена следующей главе, вместе с остальными between-моделями.

При построении прогноза нам известно планируемое количество копий первой недели релиза фильмов. Дальнейшая динамика количества копий фильмов поддается прогнозированию.

Прогноз количества копий фильма проводился с помощью пяти независимых cross-section уравнений:

, где ***t =2,3,4,5,6***;

- количество копий фильма ***i*** на неделе ***t***;

**–** среднее количество кассовых сборов, приходившееся на одну копию фильма на прошлой неделе проката (прокси доходности с одной копии фильма).

Результаты оценки данных уравнений приведены в Приложении в [Таблице 4]. Как уже упоминалось ранее, договоренности между дистрибьюторами и кинотеатрами о предоставлении копий фильмов заключаются сроком на две недели проката, поэтому количество копий второй недели проката фильма равно количеству копий первой недели – об этом говорит уравнение в первом столбце [Таблицы 4]. На третьей и всех последующих неделях проката кинотеатры, исходя из прошлой доходности фильма и прошлого количества копий фильма, оценивают перспективы повторной покупки копий данного фильма.

Стоит отметить, что динамика индекса цен, количества копий и объемов рекламы связаны со стороной «предложения» рынка кинопроката, то есть с решениями кинотеатров, дистрибьюторов и их взаимоотношениями. Поскольку в данной работе в основном использовались открытые данные о кинорынке России, то данных о стороне «предложения» у нас практически не было. В целом, описание динамики данных показателей не было успешным. Как следствие, пострадает и качество прогноза этих показателей. Для улучшения качества прогноза, необходимы более детальные данные о «внутренней кухне» киноиндустрии. А для этого необходимо участие (и внутренняя заинтересованность) игроков рынка в подобных работах.

**2.5 Модели индивидуальных отличий между фильмами**

По итогам оценки динамической модели мы получили оценки ненаблюдаемых случайной компоненты константы и тренда , определенные на выборке из генеральной совокупности фильмов, вышедших в период с 2008 по 2012 год.

У нас есть пользовательский рейтинг фильма, средняя цена билета фильма и количество копий первой недели , определенные на всей генеральной совокупности фильмов вышедших в период с 2008 по 2012 год.

Мы должны уметь предсказывать ***, ,*** для построения краткосрочных (на 6-и недельную перспективу) прогнозов кассовых сборов фильмов. Для построения долгосрочных прогнозов (свыше 6 недель) мы дополнительно должны уметь предсказывать количество копий первой недели . В данной работе долгосрочные прогнозы не будут строиться, но для полноты картины модель для количества копий первой недели будет оценена.

Средняя цена фильма разделим на среднюю инфлированную цену билета первой недели проката фильма. Таким образом, мы получим индекс цен, аналогичный использованному в динамической модели индексу.

Моделировать будем логарифмическую трансформацию индекса цен и логарифмическую трансформацию количества копий первой недели проката.

Нам необходимо построить точечные оценки, для

* – логарифма количества копий первой недели в прокате,
* - логарифма индекса цены фильма,
* – оценки случайной компоненты константы из динамической модели кассовых сборов фильма,
* – оценки случайной компоненты тренда из динамической модели кассовых сборов фильма,
* – зрительской оценки «качества» фильма.

В качестве объясняющих переменных использовались:

1. **Жанр фильма**, фильм может быть классифицирован одновременно к нескольким жанрам – группа из 23 дамми переменных
2. **Страна производства фильма**,ISO коды всех стран участвовавших в производстве картины **–** группа из 63 дамми переменных
3. **Взаимосвязи с другими фильмами**, является ли фильм продолжением, переделкой или пародией на другой фильм, содержит ли ссылки на другие фильмы – 8 количественных переменных
4. **Длительность фильма (или длинна 35мм носителя)** – одна численная переменная
5. **Технические параметры фильма**, является ли фильм широкоформатным, на каком оборудовании производилась сьемка – 4 дамми переменные
6. **Кинолаборатории задействованные в производстве фильма** – 16 дамми переменных
7. **Дистрибьютор фильма в России –** 91 дамми переменная
8. **Дата премьеры фильма в России –** 1 временная переменная
9. **Дата премьеры фильма в США** – 1 временная переменная
10. **Форматы на которых фильм вышел в Российский прокат**, 35мм/цифровой/3D **–** 3 дамми переменные
11. **Форматы на которых фильм был выпущен**, 35мм/цифровой/3D – 3 дамми переменные

При таком количестве объясняющих переменных ручная настройка моделей не возможна, поэтому анализ осуществлялся с помощью регрессионных алгоритмов адаптированных для данных высокой размерности.

Значительная часть объясняющих переменных, состояла из дамми переменных.

На первом этапе анализа, из полного списка объясняющих переменных были исключены дамми переменные, одно из значений которых соответствовало малой группе фильмов.

На втором этапе были оценены регрессионные модели по всем зависимым переменным методом LARS (Least-Angle Regression), так были отфильтрованы пять списков «потенциально» важных объясняющих переменных. Полученные списки были объединены в один общий список «потенциально» важных переменных.

Детальное описание метода LARS можно найти в статье авторов данного подхода [14]. Данный метод относится к классу алгоритмов автоматизированного построения моделей и идейно схож процедуре прямого включения (Forward Selection). Оба подхода с помощью итерационного алгоритма наполняют список переменных для модели. Результат оценки итоговой модели в обоих подходах эквивалентен обычной МНК оценке модели, сделанной по финальному списку переменных. Принципиально данные два подхода отличаются в обработке групп сильно скоррелированных переменных: процедура прямого включения обычно выбирает одну переменную из всей группы; процедура LARS выбирает все переменные данной группы. В результате, финальный список переменных модели, сформированный с помощью метода LARS, обычно содержит большее количество переменных, что может считаться преимуществом при формировании наиболее полного списка «потенциально» важных объясняющих переменных.

На последнем этапе, процедурой обратного исключения (Backward Elimination) из списка «потенциально» важных были отобраны статистически значимые переменные.

Результаты оценки зависимостей приведены в Приложении на [Таблице 5].

В общих словах сформулируем результаты оценки моделей индивидуальных отличий между фильмами:

* Любые отличия в датах релизов фильмов негативно сказываются на количестве копий;
* На протяжении наблюдаемых 5 лет наблюдался макро тренд увеличения количества копий фильмов, одновременно с сокращением цены и сокращением кассовых сборов фильмов
* Длительность кинофильма положительно связана со зрительской оценкой фильма, ценой и количеством копий фильма. Скорость падения кассовых сборов ниже у длинных фильмов
* Возрастные ограничения увеличивают стоимость билета, фильмы с ограниченной аудиторией быстрее теряют кассовые сборы, фильмы на общую аудиторию ниже оцениваются зрителями
* Жанр фильма очень существенно влияет на характеристики фильма:

1. боевики, комедии, семейные фильмы и фильмы ужасов получают преимущества в количестве копий, драма штрафуется по количеству копий
2. семейные фильмы значительно дешевле
3. по уровню кассовых сборов выигрывает комедия, фантастика и фильмы ужасов, анимация проигрывает по уровню кассовых сборов
4. зрителями лучше оценивается анимация, хуже всего зрители оценивают фильмы ужасов

* Упоминания других фильмов в картине положительно оцениваются зрителями, пародии на другие фильмы негативно влияют на оценку зрителями
* Форматы релиза фильма в России положительно влияют на количество копий фильма; 3D фильмы дороже; цифровые и 3D фильмы дольше держатся в прокате
* Формат оригинала связан только с пользовательской оценкой, классические 35мм фильмы и цифровые фильмы оцениваются лучше
* В целом технические особенности съемки и лаборатории оказались значимыми прокси качества «картинки» и спец. эффектов фильма
* Фильмы французского производства ниже по уровню кассовых сборов, но дольше живут в прокате; фильмы из Великобритании лучше оцениваются зрителями и получают преимущество в копиях; фильмы Российского производства имеют значительное преимущество в копиях и дольше держатся в прокате, но ниже оцениваются зрителями и ниже по уровню кассовых сборов
* Дистрибьютор фильма является локальным представителем одной или нескольких компаний производителей фильмов, а от производителя напрямую зависят индивидуальные характеристики выпускаемых фильмов. Мы не будем перечислять все значимые отличия между фильмами различных дистрибьюторов, но отметим очень сильную связь между дистрибьюторами и количеством копий фильмов. Данная связь может быть отражением «переговорной силы» дистрибьюторов в диалоге с кинотеатрами, например: представители востребованных производителей фильмов могут навязывать кинотеатрам бОльшее количество копий для своих посредственных по качеству фильмов, а кинотеатры вынуждены идти на уступки – либо в противном случае не получат достаточного количества копий больших премьер в будущем.

**Глава 3. Прогноз кассовых сборов фильмов**

Данная глава будет посвящена прогнозу кассовых сборов фильмов. Прогноз кассовых сборов производился в два этапа.

На первом этапе был сделан опережающий прогноз динамики кассовых сборов фильмов, результаты первого этапа описаны в начале главы.

Далее сформулирована и описана статическая модель с корректировкой Хекмана общих кассовых сборов фильмов, так мы учтем несовершенство данных использованных при анализе динамики кассовых сборов.

В конце главы приводятся результаты второго этапа опережающего прогноза кассовых сборов фильмов.

**3.1 Первый этап прогноза кассовых сборов фильмов**

**Динамический прогноз кассовых сборов фильмов**

Перед проверкой прогноза, необходимо определиться с тем какой уровень точности прогноза считать хорошим, а какой плохим применительно к задаче прогноза кассовых сборов фильмов.

Большинство авторов научных работ на тему кассовых сборов фильмов ограничиваются указанием качества линейной аппроксимации моделей, либо не указывают точность прогноза либо предпочитают в принципе не делать реального прогноза. Так например:

* в работе [2] анализ ранних (за месяц) данных мониторинга сайта Википедии позволил достичь значения коэффициента детерминации модели общих кассовых сборов ;
* анализ отзывов критиков [3] позволяет добиться для модели общих кассовых сборов;
* анализ сценария [4] позволяет добиться для модели общих кассовых сборов;
* в работе [13] достигнут коэффициент детерминации для модели кассовых сборов первой недели;
* в работе [5] можно найти значения коэффициента детерминации промежуточных моделей кассовых сборов первой недели.

Все перечисленные меры качества линейной аппроксимации взяты из моделей кассовых сборов в логарифмах. Предполагая структуру Российских данных о кассовых сборах, то модели с подобным уровнем аппроксимации имели бы ошибки лог-модели 0.38 / 0.73 / 0.99 для работ об общих кассовых сборов фильма и 0.63 / 0.90 для работ о кассовых сборах первой недели, соответственно.

Зная ошибки предложенных моделей, мы можем сделать вывод, что ошибки прогноза будут превосходить соответствующие ошибки моделей, но будут схожи им по масштабу.

Отдельно стоит отметить статью [1], в ней не приводится качество линейной аппроксимации модели, поскольку в данной работе в принципе не использовались линейные меры точности прогноза. В данной работе прогноз кассовых сборов осуществлялся с помощью обученной нейронной сети, которая классифицировала фильм в одну из групп – сформированных по шкале кассовых сборов. В работе приводится шкала кассовых сборов, по которой были сформированы группы, и приводится полная матрица попаданий/ промахов прогноза. Таких данных достаточно чтобы грубо оценить ошибку прогноза нейронной сети 0.8 для логарифма общих кассовых сборов фильма – преодоление данного уровня ошибки прогноза будем считать нашей целью.

В качестве периода для сверки прогноза были взяты шесть недель проката с 21 февраля 2013 по 3 апреля 2013 года. Прогноза строился исключительно по данным собранным до начала анализируемого периода (в период с 14 февраля по 20 февраля). В анализируемом периоде вышло в прокат 73 фильма, из которых только 17 продержались в списке двадцати самых успешных более двух недель.

В Приложении на [Диаграмме 6] приводятся результаты оценки точности прогноза динамической модели кассовых сборов, также приводятся оценками точности прогноза количества копий и индекса цен. Из диаграммы видно что точность прогноза количества копий стремительно падает от недели к неделе, что исключает возможность долгосрочных прогнозов – по крайней мере при текущем качестве априорных данных о предстоящих релизах фильмов. Точность прогноза индекса цен стабильна и находится на приемлемом уровне. Точность прогнозов кассовых сборов недели удовлетворительна и находится на уровне 0.9 (что соответствует точности 0.65 прогноза общей суммы сборов фильма), для фильмов, оказавшихся в действительности «успешными». Но априорно мы не знаем какие фильмы окажутся успешными.

Точность прогноза кассовых сборов недели для всех фильмов находится на не удовлетворительном уровне 1.1. Динамический прогноз кассовых сборов всех фильмов по общей сумме спрогнозированных кассовых сборов превысил реальную сумму кассовых сборов на 47%. Столь низкое качество прогноза связано с уже упомянутым ранее «смешением самоотбора». Наша динамическая модель, обученная на данных «успешных» фильмов, не знает о существовании «не успешных» фильмов, поэтому при прогнозе значительно преувеличивает их кассовые сборы.

Для учета эффекта «смешением самоотбора». Была построена дополнительная модель.

**3.2 Модель общих кассовых сборов фильмов**

**Статическая модель с учетом эффекта «смешения самоотбора»**

Суммарные кассовые сборы фильмов известны по всей генеральной совокупности фильмов вышедших в период с 2008 по 2012 год. Нам детально известны динамические процессы, протекающие при формировании кассовых сборов «лучших» фильмов. Мы можем адаптировать переменные динамической модели кассовых сборов для использования в статической модели - модели суммы кассовых сборов фильмов.

Переформулируем динамическую модель кассовых сборов в модель для суммы кассовых сборов «успешных» фильма:

,где – сумма кассовых сборов фильмов; и оценки случайных компонент кассовых сборов из динамической модели;

**=, =,**

**=** – все динамические переменные взвешены на среднюю скорость падения кассовых сборов фильмов .

Для учета эффекта «смешением самоотбора» была произведена коррекция Хекмана, добавлена бинарная модель описывающая селекцию фильмов:

,где –оценки по моделям индивидуальных отличий между фильмами (используем оценки, поскольку значения и случайных компонент из динамической модели определены только для «успешных»[[1]](#footnote-1) фильмов);

**–** сборы худшего из «успешных» фильмов на неделе премьеры исследуемого фильма, данная переменная несет смысл «высоты барьера» который должен преодолеть фильм для попадания в список из двадцати лучших по сборам в неделю своего релиза.

Поскольку мы предполагаем, что наша динамическая модель корректно описывала процесс формирования кассовых сборов, а критерием для фильтра фильмов на «успешные/не успешные» также являлись кассовые сборы, то все переменные модели кассовых сборов тестировались на значимость в модели селекции фильмов.

Результаты оценки модели суммарных кассовых сборов фильма приводятся в Приложении на [Таблице 6]. Все параметры динамической модели кассовых сборов значимы, в том числе и случайные компоненты кассовых сборов.

В модели селекции оказались не значимыми реклама, конкуренция, влияние праздников и разница в датах релизов фильмов. Оценки случайных компонент константы и тренда, количество копий фильма и цена фильма значимо влияют на вероятность попадания фильма в список успешных. Чем выше барьер сборов который необходимо преодолеть фильмам для попадания в список из двадцати лучших по сборам на данной неделе, тем ниже шансы стать «успешными» у фильмов, вышедших в прокат на данной неделе.

Значимо отрицательно на вероятность попадания в список успешных влияет одна техническая особенность съемки фильма: если съемка фильма осуществлялась на узкую пленку, а потом технически копировалась на стандартную 35мм – к данной процедуре обычно прибегают малобюджетные фильмы.

Фильмы двух дистрибьюторов: PF- Premium Film и VLG- Вольга имеют высокую вероятность оказаться в списке «успешных». Отдельно стоит отметить, что оценка влияния дамми некоторых дистрибьюторов оказалось невозможной в модели селекции. Эти дистрибьюторы приведены отдельно в [Таблице 7]. У каждого из перечисленных дистрибьюторов фильмы были либо все «успешные», либо все «не успешные» на анализируемом периоде данных. К прогнозу вероятности «успеха» новых фильмов данных дистрибьюторов нужно подходить с особым вниманием.

**3.3 Второй этап прогноза кассовых сборов фильмов**

**Статический прогноз суммы кассовых сборов фильмов**

На первом этапе прогноза кассовых сборов фильмов нами был построен прогноз кассовых сборов недели для фильмов, данный прогноз мы используем для построения грубого списка из двадцати «лучших» фильмов. Мы возьмем прогноз кассовых сборов самого плохого фильма из грубого списка «лучших» фильмов и будем использовать его в качестве оценки .

Динамический прогноз количества копий фильмов (по неделям) из первого этапа прогнозирования мы используем для оценки конкуренции **.**

Остальные переменные модели суммы кассовых сборов нам либо известны, либо мы их прогнозируем при помощи моделей индивидуальных отличий между фильмами.

Результаты прогноза суммы кассовых сборов фильмов:

* точность классификации фильмов на «успешные/не успешные» 86%
* точность прогноза суммы кассовых сборов фильмов 0.76, для фильмов предсказанных как успешные (для сравнения, точность прогноза суммы кассовых сборов динамической модели для этих фильмов 1.02)
* точность прогноза суммы кассовых сборов фильмов 0.55, для фильмов в действительности оказавшихся успешными (для сравнения, точность прогноза суммы кассовых сборов динамической модели для этих фильмов 0.65)

Введя корректировку на эффект «смещения самоотбора» мы не только улучшили качество прогноза модели, но и научились с высокой вероятностью предсказывать «успешные» фильмы, то есть фильмы более 2 недель продержавшиеся в списке двадцати лучших по сборам.

**Заключение**

В данной работе была детально исследована динамика кассовых сборов фильмов в Российском прокате. Обнаружено значимое влияние таких динамических аспектов как конкуренция между фильмами в прокате, влияние праздников и реклама дистрибьюторов фильмов. Также было найдено, что скорости падения кассовых сборов фильмов различны, различия значимы и частично объясняются характеристиками фильмов.

В работе был предложен двух шаговый алгоритм краткосрочного (на шесть недель) прогноза кассовых сборов «успешных»[[2]](#footnote-2) фильмов: на первом шаге, строится грубый динамический прогноз кассовых сборов всех фильмов из списка предстоящих премьер; на втором шаге, предсказывается вероятность попадания фильма в список «успешных» и предсказывается суммарные кассовые сборы «успешных» фильмов.

Предложенный алгоритм продемонстрировал высокую точность (86%) предсказания попадания фильмов в список «лучших». Точность прогноза суммы кассовых сборов «успешных» фильмов в России у предложенного алгоритма 0.76) находится на одном уровне с альтернативными подходами прогноза кассовых сборов фильмов в США 0.8).

Точность прогноза кассовых сборов фильмов в России может быть улучшена в двух различных направлениях:

* распространение полученных подходов на все фильмы в Российском прокате – для этого необходимы дополнительные данные о динамике кассовых сборов всех фильмов в Российском прокате (а не только двадцати лучших за неделю);
* улучшение прогноза характеристик «предложения» киноиндустрии: прогноза цен билетов и прогноза количества копий – для этого необходимы дополнительные данные о внутренней структуре кинорынка: о дистрибьюторах, кинотеатрах и их взаимоотношениях.

Отдельно стоит отметить, что в целом точность прогнозов в киноиндустрии достаточно низкая. Объясняется это тем, что объекты исследования (фильмы) крайне неоднородны, а свойства этих объектов плохо формализуются статистически. Однако данное исследование является попыткой продвинуться в этой области.

**Список использованной литературы, источники данных**

1. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications 30: 243-254.
2. Marton M, Taha Y, Janos K (2012) Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data. [[arXiv: <http://arxiv.org/abs/1211.0970>](http://arxiv.org/abs/1211.0970)].
3. Kennedy A (2008) Predicting Box Office Success: Do Critical Reviews Really Matter? Berkeley Projects.
4. Eliashberg J, Hui S K and Zhang Z J (2007) From Storyline to Box Office: A New Approach for Green-Lighting Movie Scripts. Management Science 53(6), 881–893.
5. Antipov E, Pokryshevskaya E (2010) Accounting for Latent Classes in Movie Box Office Modeling. [SSRN: <http://ssrn.com/abstract=1729631>]
6. Achal A, Manas P (2011) User rating prediction for movies [<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.122.4298>]
7. [Andrei O, Mathias B, Manos T , Maarten R (2012) Predicting IMDB movie ratings using social media. Proceedings of the 34th European conference on Advances in Information Retrieval, Barcelona, Spain](http://dl.acm.org/citation.cfm?id=2260703&CFID=330766070&CFTOKEN=32126914)
8. Производственный календарь с 2008 по 2012г / РФ
9. The Internet Movie Database (IMDb). [<http://www.imdb.com/>]
10. Бюллетень Кинопрокатчика. [[http://www.kinometro.ru](http://www.kinometro.ru/)]
11. Erik B, Kjersti-Gro L, Terje S (2000) Heterogenety in return to scale: a random coefficient analysis with unbalanced panel data. Discussion Paper No.292, December 2000 Statistics Norway, Research Department [<http://www.ssb.no/a/publikasjoner/pdf/DP/dp292.pdf>]
12. Markus M (2010) Unbalanced Panel Data Models, Baltagi Textbook - Chapter 9. June 2010, Department of Economics University of Vienna [<http://homepage.univie.ac.at/robert.kunst/pan2010_pres_mayer.pdf>]
13. Joshi M, Das D, Gimpel K, Smith N (2010) Movie reviews and revenues: An experiment in text regression. In: Proceedings of NAACL-HLT 2010, Short Papers Track [<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.184.2470>]
14. Efron B, Hastie T, Johnstone I and Tibshirani R(2004) Least Angle Regression. Annals of Statistics 32 (2): pp. 407–499 [<http://arxiv.org/pdf/math/0406456.pdf>]

**Приложение**

Диаграмма 1. Динамика первых 4 недель кассовых сборов фильмов крупнейших дистрибьюторов. За 100% взяты сборы первой недели фильма. На 2,3 и 4 неделях представлены средние значения долей от сборов первой недели по всем фильмам дистрибьютора.

Диаграмма 2. Макро-динамика цен билетов в кинотеатрах России  
 за период с 2008 по 2012 год.

**Средняя годовая инфляция цен на билет 10.3%**

Диаграмма 3. Зависимость индекса кассовых сборов от количества копий.



Диаграмма 4. Внутри- недельная сезонность и влияние праздников. За 100% взяты кассовые сборы типичной недели без праздников - темные ячейки демонстрируют сезонность внутри такой недели. Белыми ячейками обозначена надбавка к сборам недели при наличии праздника в данный день недели. Если на неделе было несколько праздников, то надбавки аддитивно суммируются.



Диаграмма 5. Латентный показатель конкуренции между фильмами.

Таблица 1. Сводные статистики спецификаций модели:

1) Оценка максимального правдоподобия модели со случайным эффектом; 2) Модель со случайными коэффициентами – со случайной константой ; 3) Модель со случайными коэффициентами – со случайными независимыми константой и трендом ; 4) Модель со случайными коэффициентами – со случайными константой , трендом и корреляцией между ними.



Таблица 2. Результаты оценки случайных параметров: константы , тренда и корреляции между ними, а также случайной ошибки модели.



Таблица 3. Результаты оценки фиксированного влияния параметров модели со случайными коэффициентами.



Таблица 4. Зависимость количества копий фильма от количества копий прошлой недели и средней доходности копии прошлой недели. Первый столбец соответствует уравнению для второй недели проката, второй столбец соответствует уравнению третьей недели и т.д.



Таблица 5. Модели индивидуальных отличий между фильмами.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |
| Отставание в датах релиза в России | | -0.247\*\*\* |  |  |  |  |
| Опережение в датах релиза в России | | -0.234\*\*\* |  |  |  |  |
| Макро тренд, номер недели | | 0.002\*\* | -0.001\*\*\* |  | -0.002\*\*\* |  |
| Длительность фильма, мин | | 0.004\* | 0.002\*\*\* | 0.002\*\*\* |  | 0.013\*\*\* |
| Ограничения аудитории | Parental Guidance Suggested |  |  |  |  | -0.360\*\*\* |
| Parents Strongly Cautioned | 0.209\*\* |  | -0.042\*\* |  | -0.230\*\*\* |
| Restricted (18+) |  | 0.032\*\* | -0.035\* |  |  |
| Жанр | Action | 0.417\*\*\* |  | -0.028\* |  |  |
| Adventures | 0.210\* |  |  |  |  |
| Animation | 0.249\* |  | 0.054\* | -0.118\* | 0.515\*\*\* |
| Comedy | 0.309\*\*\* | 0.030\*\* | 0.030\* | 0.104\*\*\* | -0.213\*\*\* |
| Drama | -0.198\*\* | 0.022\* |  |  | 0.250\*\*\* |
| Family | 0.431\*\*\* | -0.154\*\*\* |  |  |  |
| Fantasy |  |  |  | 0.117\*\* |  |
| Horror | 0.400\*\*\* |  |  | 0.145\*\* | -0.548\*\*\* |
| Romance | 0.150\* |  |  |  |  |
| Thriller | 0.243\*\*\* |  |  |  |  |
| Связи с другими фильмами | Упоминания других фильмов |  |  |  |  | 0.014\*\*\* |
| Пародии на другие фильмы |  |  |  |  | -0.068\*\*\* |
| Форматы релиза в России | 35мм | 0.231\*\* |  |  |  |  |
| цифровое | 0.566\*\*\* |  | 0.033\*\* |  |  |
| 3D | 0.691\*\*\* | 0.247\*\*\* | 0.062\*\*\* |  |  |
| Форматы оригинала | 35мм |  |  |  |  | 0.172\*\* |
| цифровое |  |  |  |  | 0.134\* |
| Технические особенности съемки | Съемка осуществлялась на узкую пленку, а потом технически копировалась на стандартную 35мм |  |  |  |  | 0.369\*\* |
| При съемке использовались сферические линзы |  |  |  |  | 0.235\*\* |
| При съемке использовались анаморфные линзы | 0.144\* |  |  |  |  |
| отсутствовала техническая спецификация | -0.230\*\* | -0.056\*\*\* |  | -0.140\*\* |  |
| Лаборатория | CinePostproduction | -0.722\* | 0.106\* |  |  |  |
| DeLuxe |  |  |  | 0.079\* |  |
| DeLuxe, UK | -0.441\*\* | 0.064\* |  | -0.196\* |  |
| FotoKem Digital Film Services |  |  | 0.064\* |  |  |
| Imagica Corporation |  | 0.120\* |  |  |  |
| Laboratoires LTC |  | 0.074\* |  | 0.270\* | 0.489\*\* |
| Laser Pacific Labs |  |  |  |  | -0.688\*\* |
| Technicolor, NewY,Holly,LosA,Lond |  |  |  |  | 0.246\*\* |
| Страна производитель | Франция |  |  | 0.067\*\*\* | -0.117\* |  |
| Великобритания | 0.171\* |  |  |  | 0.210\*\* |
| Россия | 0.299\*\* |  | 0.065\*\* | -0.120\* | -0.402\*\*\* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |
| Дистрибьютор в России | BVSPR- Buena Vista Sony Pictures Releasing | 0.866\*\*\* |  |  |  |  |
| BZL- Bazelevs |  |  |  | 0.535\* |  |
| CAE- Каскад фильм | 0.991\*\*\* |  |  |  |  |
| CAO- Каро Премьер | 1.672\*\*\* |  | -0.067\*\*\* | 0.146\*\* | -0.238\* |
| CDDC- Каравелла DDC |  | -0.106\*\*\* |  |  |  |
| CMF- Кармен Фильм | -2.323\*\*\* | -0.147\*\*\* |  |  |  |
| CP- Централ Партнершип | 1.603\*\*\* |  | -0.083\*\*\* |  | -0.475\*\*\* |
| CPCL- CP Classic | -0.441\* | 0.122\*\*\* |  |  |  |
| CPP- Централ Партнершип Paramount | 1.610\*\*\* |  |  |  |  |
| CPRG- Cinema Prestige | -1.084\*\*\* |  | 0.292\*\* |  | -0.560\* |
| CRP- Каропрокат | 1.752\*\*\* |  | -0.118\*\*\* | 0.321\*\*\* | -0.668\*\*\* |
| CWF- Кино без границ | -1.226\*\*\* |  |  |  | 0.471\*\*\* |
| FOX- Fox | 1.529\*\*\* |  |  |  |  |
| LEA- Леопарт | -1.200\*\*\* |  |  |  |  |
| LEO- Леополис |  |  | -0.286\*\*\* |  |  |
| LUX- Люксор | 1.524\*\*\* |  | -0.109\*\*\* |  | -0.397\*\* |
| NKI- Наше кино | 0.951\*\*\* | 0.065\* | -0.164\*\*\* |  |  |
| NKI- Наше кино/CRP- Каропрокат |  |  |  | -0.879\* |  |
| P&I- P&I Films | -1.846\*\*\* |  |  | -1.435\*\*\* |  |
| PF- Premium Film | -1.420\*\*\* | 0.092\*\*\* |  |  |  |
| PNR- Панорама Кино |  | -0.548\*\*\* |  |  | -0.720\* |
| PRD- Парадиз | 0.727\*\*\* | 0.084\*\*\* | -0.157\*\*\* |  | -0.370\*\*\* |
| PRM- Компания Пирамида | 1.465\*\*\* | -0.222\*\*\* |  |  | -0.511\* |
| RCC- RUSCICO | -1.035\*\*\* | -0.156\*\*\* |  |  |  |
| RUR- Русский Репортаж | -1.607\*\*\* |  |  |  |  |
| TFD- Top Film Distribution | 1.190\*\*\* |  |  |  |  |
| UPI- UPI | 1.519\*\*\* |  | -0.094\*\*\* | 0.242\*\*\* |  |
| VLG- Вольга | 0.463\*\* | 0.057\* |  |  | 0.508\*\*\* |
| WDSSPR- WDSSPR | 1.321\*\*\* |  |  | 0.266\*\*\* |  |
| WST- West | 1.343\*\*\* |  |  | 0.207\*\* | -0.348\*\* |
|  | \_cons | 2.667\*\*\* | -0.152\*\*\* | -0.174\*\*\* | 0.201\*\*\* | 4.679\*\*\* |
|  |  |  |  |  |  |  |
|  | N | **1463** | **1 508** | **695** | **695** | **1499** |
|  | r2\_a | **0.788** | **0.329** | **0.246** | **0.286** | **0.322** |
|  |  |  |  |  |  |  |
|  |  | legend: \* p<0.05; \*\* p<0.01; \*\*\* p<0.001 | | | | |

Диаграмма 6. Точность прогноза динамической модели кассовых сборов и точность прогноза динамики объясняющих переменных данной модели.

Таблица 6. Результаты оценки модели общих кассовых сборов фильмов.



Таблица 7. Дистрибьюторы, все фильмы которых в анализированном периоде попадали в одну группу селекции.

|  |  |
| --- | --- |
| Дистрибьюторы, фильмы которых вышедшие с 2008 по 2012год были все «успешными» | Дистрибьюторы, фильмы которых вышедшие с 2008 по 2012год были все «не успешными» |
| |  | | --- | | BZL- Bazelevs | | NKI- Наше кино/CRP- Каропрокат | | |  | | --- | | CMF- Кармен Фильм | | LEA- Леопарт | | PNR- Панорама Кино | | PRM- Компания Пирамида | | RCC- RUSCICO | |

1. «успешными» фильмами считались фильмы более 2 недель продержавшиеся в списке двадцати лучших по сборам [↑](#footnote-ref-1)
2. «успешными» фильмами считались фильмы более 2 недель продержавшиеся в списке двадцати лучших по сборам [↑](#footnote-ref-2)