МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ Государственное автономНОЕ образовательное

учреждение высшего образования

«Новосибирский НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ государственный университет»   
(нОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет экономический

Кафедра применения математических методов в экономике

Направление подготовки 38.03.01 Экономика

Уровень подготовки бакалавриат

Группа 19701.2

1. **КУРСОВАЯ РАБОТА**
2. Балабаев Владислав Дмитриевич
3. (Фамилия, Имя, Отчество автора)

Тема работы Эмпирическая оценка спроса на товарную категорию на примере данных пермской торговой сети

**Научный руководитель**

к.э.н., профессор,

кафедра "применения математических методов в экономике и планировании" ЭФ

Шильцин Евгений Александрович

(Фамилия, И., О.)

Новосибирск

2022

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 3](#_Toc103968100)

[ГЛАВА 1. ХАРАКТЕРИСТИКА ИМЕЮЩИХСЯ ДАННЫХ И ИХ АНАЛИЗ 5](#_Toc103968101)

[1.1. Описание данных пермской торговой сети «Семья» по продаже соков 5](#_Toc103968102)

[1.2. Анализ данных и создание массива из переменных 6](#_Toc103968103)

[ГЛАВА 2. КРАТКАЯ ХАРАКТЕРИСТИКА ПРИМЕНЯЕМЫХ МОДЕЛЕЙ ДЛЯ ОЦЕНКИ СПРОСА 13](#_Toc103968104)

[2.1. Эконометрические модели, основанные на распределении Пуассона 13](#_Toc103968105)

[2.2. Алгоритмы машинного обучения, в основании которых лежит алгоритм регрессионного дерева решений 17](#_Toc103968106)

[2.3. Формирование алгоритма долгосрочного прогнозирования 21](#_Toc103968107)

[ГЛАВА 3. ОПИСАНИЕ И ОЦЕНКА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ 23](#_Toc103968108)

[3.1. Сравнение результатов рассмотренных моделей на имеющихся данных 23](#_Toc103968109)

[3.2. Несовершенство представленных моделей и неэффективность потенциальной оптимизации по ценам 29](#_Toc103968110)

[ЗАКЛЮЧЕНИЕ 31](#_Toc103968111)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 32](#_Toc103968112)

# ВВЕДЕНИЕ

**Актуальность темы работы.** Актуальной задачей розничной торговли является, с точки зрения максимизации прибыли розничного предприятия, задача оптимального ценообразования. Как следствие, подзадача - оптимальное ценообразование в рамках категории продукта, которое бы учитывало установление скидки оптимального размера в какой-то момент. Обе эти подзадачи, а также более общая задача ценообразования в розничной торговле с учетом конкуренции между сетями имеют относительно однородную структуру данных, примерно одинаковые подходы, с которых можно начать, чтобы приблизиться к решению такого рода проблемы.

В рамках задачи ценообразования рассматривается дифференцированный товар, то есть товар из определенной товарной категории или конкретного магазина определенной сети, который наблюдается на определенном временном промежутке. Имея информацию о продажах этого продукта, его цене и характеристиках, а также аналогичную информацию о его несовершенных конкурентах и некоторых внешних условиях для рынка, мы можем прогнозировать равновесный уровень продаж.

Контролируя модель продаж на предмет возможного влияния решений продавца, например, решения об установлении цены или скидки, мы также можем изолировать функцию спроса в модели равновесных продаж. Правильная оценка функции спроса в конечном итоге позволит, зная реакцию спроса на цену, установить оптимальный вектор цен, который может содержать в себе временные промежутки с наличием скидок, для продавца, который позволит максимизировать прибыль для конкретного товара или категории товаров.

Как следствие, необходимой базой для решения подобной проблемы выступает задача осуществления правдоподобного долгосрочного прогноза, при котором, зная реакцию спроса на значения параметров цен, можно, при использовании подходящего оптимизационного итеративного метода, максимизировать значение прибыли по соответствующей товарной категории на прогнозируемом промежутке времени.

**Целью** предлагаемой курсовой работы на третьем курсе, которая будет последовательно развиваться в рамках ВКР, является формирование долгосрочного прогноза равновесных продаж по товарной категории субъекта розничной торговли на основе эмпирической оценки спроса на товары данной категории.

Определение поставленной цели позволяет выделить определенный перечень задач:

1. Определить сегмент розничной торговли, а также товарную категорию
2. Провести описательный анализ имеющихся по товарной категории данных, определив характерные особенности
3. Рассмотреть подходы, которые позволят провести эффективную оценку имеющихся данных
4. Осуществить эмпирический анализ имеющихся данных на основе предложенных моделей и провести оценку полученных результатов

**Объект исследования**: Спрос на товары розничных торговли

**Предмет исследования**: Формирование долгосрочного прогноза равновесных продаж по товарной категории субъекта розничной торговли на основе эмпирической оценки спроса на товары данной категории.

# ГЛАВА 1. ХАРАКТЕРИСТИКА ИМЕЮЩИХСЯ ДАННЫХ И ИХ АНАЛИЗ

## 1.1. Описание данных пермской торговой сети «Семья» по продаже соков

Данные, на которых основывается настоящая курсовая работа представлены Пермской торговой сетью “Семья” по продаже соков. Данная товарная категория представлена во всех магазинах сети, обладает относительно стабильными продажами в течение года и является относительно независимой от продаж в других товарных категориях.

Единицей товара в данных является SKU (Stock Keeping Unit) – единица складского учета, уникальный код, используемый в торговле для определения конкретной единицы товара в магазине или на складе. Такой формат необходим для того, чтобы товар можно было легко найти или отследить.

Из информации, содержащейся в названии SKU, мы можем вытащить объем упаковки, производителя, вкус сока, процент содержания натурального сока, материал изготовления пачки и содержание в соке различных добавок. Данные параметры SKU образуют набор характеристик сока, которые не меняются во времени.

Единицей наблюдения является вектор-строка, содержащая данные о том, в какой день, в каком магазине и по какой цене была совершена продажа определенного количества конкретного SKU.

Проблемой является то, что в чеках отдельно не хранится цена продажи без учета скидки конкретному покупателю. Поэтому если в каком-то магазине были нулевые продажи, по нему невозможно посчитать цену. Однако, цены во всех магазинах сети в один и тот же день являются одинаковыми, поэтому если хотя бы в одном магазине данный SKU продавался, то выручка от SKU во всех магазинах в день, деленная на общий объем, покажет довольную точную оценку цены. Однако возникают случаи, когда ни в одном магазине сок не был продан в течение дня, тогда и цена и средняя цена по всем магазинам будут пустыми, а общее количество продаж по нему будет нулевое.

Относительный показатель количества пропущенных наблюдений по ценам к общему количеству наблюдений составляет 50,7%, что равносильно тому, что половина наблюдений соответствует нулевым продажам по поставляемому SKU в конкретный день в определенном магазине. Такой же показатель по вектору средних цен составляет лишь 7,7%, а так как аппроксимация цены является крайне точной, то явным решением является выбор вектор-столбца средних цен.

Другая проблема с ценами заключается в том, что известна только фактическая цена продажи и тот факт, что на сок была скидка, но размер скидки неизвестен. На самом деле потребитель наблюдает как зачеркнутую цену, так и новую цену, а мы видим только новую. Некоторой оценкой для зачеркнутой цены, а соответственно и величине скидки в абсолютном и относительном выражении**,** может служить цена SKU в предыдущие периоды.

Предоставленные данные содержат информацию только о тех соках, которые были в продаже. Если остатки равны нулю, то в тот день на полке не было сока, поэтому наблюдений за такими SKU нет в выборке. Притом, как было описано выше, все нулевые продажи означают, что сок фактически на полке был, но его никто не купил, поэтому такие наблюдения присутствуют.

## 1.2. Анализ данных и создание массива из переменных

Данные доступны за 5 лет, в разрезе дней, магазинов и SKU формируют около 800 тыс. наблюдений.

Однако, для уменьшения общего объема времени вычислительных процессов, занимаемого при моделировании, а также для упрощения самой задачи была избрана подвыборка, которая содержит в себе только наблюдения из магазина “Семьи” на Борчанинова - формат большого магазина в центре города. Тем самым размерность данных была сокращена до 520 тыс. наблюдений.

Так, на представляемом графике (рис. 1) можно увидеть, что характеристика распределения покупок самих категорий соков практически не меняется при сравнивании двух совершенно разных по формату супермаркетов, отличается лишь величины потоков продаж.

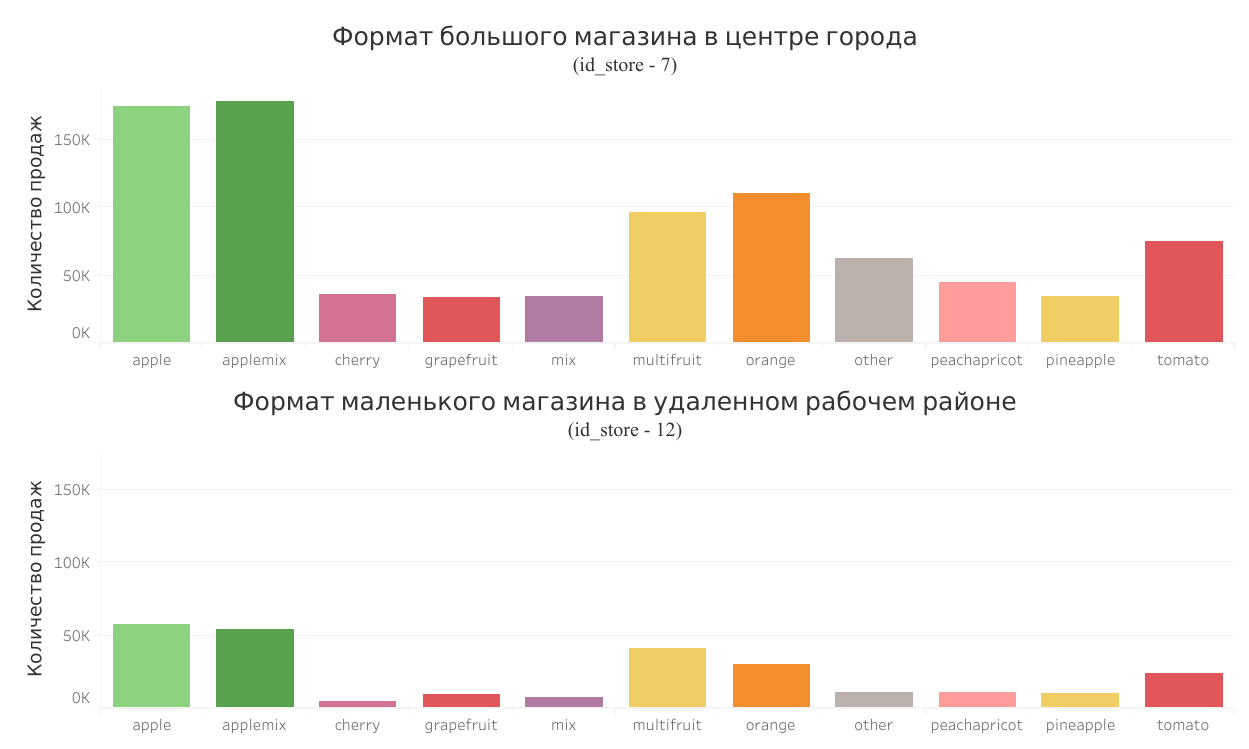


Рисунок 1 – Сопоставление разных форматов магазинов розничной сети “Семья”  
 по распределению общего количества продаж по вкусам соков за период 2009-2014 гг.

Общее число уникальных SKU составляет 463 единицы, поэтому, руководствуясь приведенными выше принципами, выполним запрос данных только соков, вкус которых – томатный. Уменьшение до настоящей подвыборки обусловлено наличием малой перекрестной эластичности с другими характеристиками вкуса данной товарной категории (яблочный микс, мультифрукт, апельсиновый, ананасовый и т.д., которые можно, группируя, обозначить за “сладкие”), тем самым, не ухудшая, а только конкретизируя будущие результаты, мы добиваемся более удобной для эмпирического анализа товарной группы. После проведения данной операции количество уникальных SKU составило 34 единицы.

Вместе с этим, ссылаясь на несовершенство имеющихся данных, отрицательные значения продаж, коих было меньше 10, были переформатированы в наблюдения с нулевыми продажами.

Графически анализируя приведенные данные видно, что у большинства SKU наличествуют продолжительные периоды полного отсутствия продаж за имеющийся период времени. Более того, существуют отдельные номера SKU, общее количество продаж которых за период в 5 лет не превышало и 50, так, например, SKU с номерами 465321 и 465424 имели 29 и 42 проданных единицы соответственно, что говорит либо о краткосрочном характере существования производства самого товара или о непродолжительном испытании товара самой торговой сетью.

Делая предположение, что такого рода товары имели недостаточное или, иначе говоря, статистически малое воздействие на другие SKU данной товарной категории, поставки которых осуществлялись относительно стабильно, а также считаясь с тем фактом, что эмпирическое модельное оценивание таких товаров было бы несостоятельным по причине малого размера выборки, осуществим усечение количества уникальных SKU по критерию того, что агрегированное количество продаж за весь период составляло больше 600, то есть приблизительно хотя бы 1 покупка в 3 дня. В результате чего после проведения данной операции количество уникальных SKU составило 25 единиц, то есть сократилось на 9 единиц.

Также изобразим распределение имеющихся данных по всем количествам продаж SKU, по всем датам наличия поставок товара. Данная иллюстрация (рис. 2) нам понадобится дальше, когда мы будем рассматривать класс подходящих моделей.

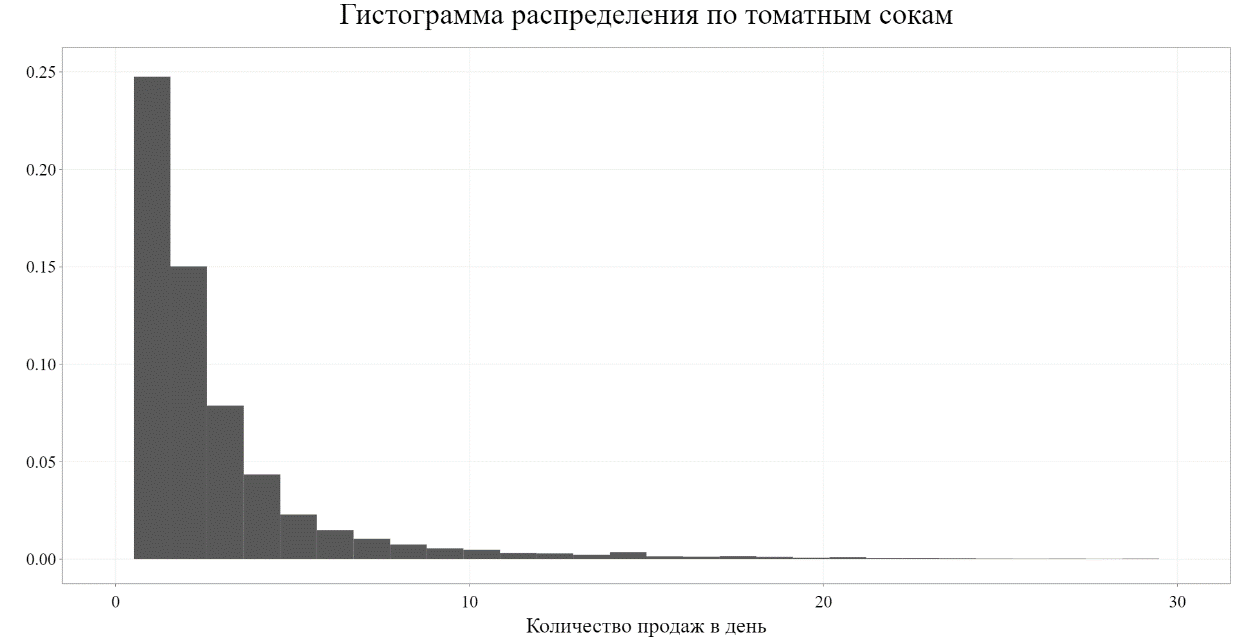


Рисунок 2 – Гистограмма распределения значения продаж за день по томатным сокам за период наличия поставок по всем SKU

Вместе с этим, изменим сам формат данных до более удобоваримого вида. Текущий формат данных представляем собой так называемый melt table, который представляется в виде длинного списка всех присутствующих наблюдений. Для того, чтобы эти данные имели матричный вид, с которым непосредственно работают библиотеки машинного обучения, переведем их в формат pivot table, где самими переносимыми значениями будут являться значения продаж и цен, как индекс возьмем дату, а в качестве группирующего элемента возьмем значения ID SKU. В результате мы получили матрицу 2191 на 50.

Добавим в приведенную выше матрицу группу факторов, которые соответствуют SKU, как сами наличия поставок, где каждый фактор будет являть собой dummy переменную, принимающую значение 0, если значение является не числовым (NaN), и 1 иначе, то есть, когда поставки осуществляются. После чего совершим замену всех пропущенных значений (NaN) на ноль. Данная операция необходима для того, чтобы модели могли учитывать сам факт значения поставок и работать со всеми наблюдениями, напротив, если бы данные не были подвергнуты текущему преобразованию, то количество строк, не содержащих пропущенные значения, было бы нулевым, что было бы критичным для моделей из библиотек машинного обучения.

Руководствуясь теми же соображениями, заменим пропущенные значения цен на медианные значения. Методом замещения выступила обычная замена на медианную, потому что, например, заполняя методом “замещения вперед” или “замещения назад” мы можем попасть на значения дней со скидкой и тем самым ее фактически продлить, что негативно повлияет на значения калькулируемой в будущем дополнительной таблицы с периодами скидок, которая станет хорошим подспорьем в визуальном анализе полученных результатов. Заполнять же пропущенные значения нулем было бы теоретически неверно, потому как эконометрические модели будут неверно строить взаимосвязь количества продаж от цены, и прямая ценовая эластичность окажется смещенной.

Также была осуществлена корректировка на инфляцию по индексу потребительских цен на фруктовые соки, Пермский край, в % к предыдущему месяцу, после чего цены на представленные SKU были в постоянных ценах 2009-01-01. Прирост ИПЦ на фруктовые соки в Пермском крае за период наличия данных составил 48%. Был взят именно ИПЦ по фруктовым сокам, потому как ИПЦ на категорию свежие томаты имеет слишком большой разброс и сильно зависит от сезонности, а ИПЦ на консервы томатные слишком завышен по сравнению со средним относительным приростом цен на сами имеющиеся соки, который в среднем по всем сокам составил приблизительно 25%, а прирост предлагаемого ИПЦ составил 63%. Более точной аппроксимирующей для инфляции на томатные соки найти не удалось.

Создадим отдельную таблицу, в которой будут находиться периоды скидок и их значения. Калькулироваться скидка будет следующим образом: если значение разницы медианного значения цены и текущего значения цены по отношению к медианному значению цены составляло меньше 20%, то бралось нулевое значение, иначе же бралось само калькулируемое значение. Стоит отметить, что 4 из 25 SKU на протяжении всего периода не имели скидок, поэтому было принято решение, упрощая дальнейшую созидательную работу, отбросить приведенные SKU. В результате перечисленных выше действий текущий массив состоит из 2191 наблюдения и 63 переменных, как 21 SKU на 3 значения – количества продаж, цены и наличия поставок.

Делая небольшое отступление, отметим, что нам интересен именно анализ воздействия изменения цены, а в частности, установление скидки, на количество продаваемых единиц SKU, так как, согласно данным, беря в расчет только периоды наличия поставок, периоды скидок генерируют в среднем в 5,0897 больше выручки, чем в периоды наличия стабильных цен, со стандартным отклонением в 4,3838. С максимальным значением в 21,4296 и минимальным в 0,5996. Рассматривался именно показатель выручки, так как значения цен поставок неизвестны, из-за чего калькуляция прибыли невозможна. А спекуляции с тем, какими цены поставок могли бы быть, если брать какой-то определенный уровень маржинальности, и впоследствии рассчитывать прибыль, не привели бы к каким-то стоящим выводам, что, к тому же, подкрепляется тем фактом, что в большинстве случаев поставщик сам часто назначает более низкую цену поставляемой партии, с условием, что на данные товары будет назначена скидка. Кроме того, потенциально возникающий аргумент, состоящий в том, что в условиях, когда в большинстве случаев сам поставщик назначает период скидок и скидочный процент, нивелируется тем фактом, что сам розничный торговец, имея доказательную базу того, в какие периоды назначение скидки является более предпочтительным, имеет возможность корректировать действия поставщика, добиваясь более высоких взаимовыгодных результатов.

Добавим категориальные переменные в виде нескольких dummy переменных, которые бы отвечали за текущий день недели и месяц наблюдения, предварительно взяв за базовые, то есть отбросив из массива, понедельник для категориальной переменной дня недели и январь для категориальной переменной месяца. В результате текущие данные имеют размерность 2191 на 80.

Реализация добавления в таблицу российских праздников, которые влияют на производственный календарь была отвергнута по нескольким причинам. Во-первых, учет предпраздничной покупки раздул бы количество dummy переменных в задаваемое количество лагов, из-за чего полученный массив мог оказаться близким к вырожденному, во-вторых, при первичном анализе учет наличия праздников почти не дал практической статистически доказуемой пользы, в-третьих, абсолютное большинство праздников мало ассоциируются с томатным соком, в-четвертых, при работе с библиотеками, содержащими в себе национальные праздники, влияющие на производственный календарь, не учитывается тот факт, что из года в год большое количество праздников меняет дату из-за того, что некоторые из них переходящие, а некоторые попадают под дату выходных.

Перейдем к формированию непосредственно самих матриц объясняющих переменных. Так как будущие модели будут сформированы в векторном виде, то для каждого SKU нужно, используя циклы, составлять свой собственный массив данных, который бы имел общую конфигурацию.

Начнем с того, что в день, в который мы хотим оценить сколько будет куплено по какому-то конкретному SKU, мы не знаем значений того, сколько было куплено по другим SKU, а значит, необходимо сместить значения продаж остальных SKU на один день. Само же значение продаж в качестве объясняемой переменной будет выступать в качестве i1, i2, i3 лагов, которые подбирались индивидуально для каждого ряда как первые три самые наибольшие значения частной автокорреляционной функции. В основном, самыми распространенными значениями служили 1 лаг, который выступал самым значимым для всех рядов, 2 и 3 лаги, однако, среди значимых содержались 4, 5, 7 и даже 8 лаги. Однако, так как модель должна учитывать непосредственно факт взаимодействия цены и количества продаж, учитывая в лагах продаж лишь значения неценового спроса, стоит воспользоваться следующим эконометрическим приемом. Рассчитывается модель, объясняющая зависимость количества продаж в день от цены и фактора наличия поставок, и потом за переменную чистого неценового спроса берется разность фактического количества продаж и оценка объясненного ценой количества продаж.

Вместе с этим, можно укрупнить чистое неценовое количество продаж за последнюю неделю (неценовой спрос за неделю) по самому таргетированному SKU и неценовое количество продаж по сумме всех остальных SKU, которое потенциально может служить хорошей статистической переменной.

Однако, мы знаем значения всех цен, наличие поставок и то, к какому дню недели и месяцу относится текущее наблюдение. Поэтому добавим к матрице переменных цену таргетированного SKU и еще i1, i2, i3 лаги, остальные текущие цены помноженные на dummy переменную наличия поставок, день, месяц, наличие поставок самого SKU, день и месяц. Также, как и прежде, можно укрупнить цены, понимая, что большая часть параметров при факторах цен остальных SKU может оказаться незначимой, тогда как сам факт наличия скидки (или скидок) может выступать значимым фактом, до переменной, которая бы отвечала за то, какое количество остальных товаров в текущий день имеет скиду и соответственно переманивает спрос.

По итогу мы получаем следующую матрицу объясняющих переменных для определенного SKU:

|  |  |
| --- | --- |
|  | (1.1) |

где – текущая цена таргетируемого SKU, а также ее лаги; – вектор текущих цен на другие SKU, помноженный на dummy переменную поставок соответствующих SKU; – лаги неценового спроса на таргетируемый SKU; – dummy переменная наличия поставок; – суммарный неценовой спрос за последние 7 дней (не включая текущий) на таргетируемый SKU; – суммарный неценовой спрос за последние 7 дней (не включая текущий) на все остальные SKU; – категориальная переменная дня недели; – категориальная переменная месяца; – количество скидок по всем товарам в текущий день; – вектор значений неценового спроса на остальные SKU в предыдущий период.

Получив матрицу объясняющих переменных, а также вектор значения продаж сока по определенному SKU, проведем разбиение на тренировочную часть и тестовую в соотношении 0,85 и 0,15. С помощью использования циклов получим приводимую выше матрицу для каждого SKU.

# ГЛАВА 2. КРАТКАЯ ХАРАКТЕРИСТИКА ПРИМЕНЯЕМЫХ МОДЕЛЕЙ ДЛЯ ОЦЕНКИ СПРОСА

## 2.1. Эконометрические модели, основанные на распределении Пуассона

Оценивая класс моделей, которые могут быть применены к имеющимся данным, прежде всего, стоит отметить, что значения количества продаж, которые имеются в данных, являются дневными, а потому, большое количество наблюдений содержат нулевое значение продаж, вместе с этим, спецификой имеющихся данных является то, что они представляют собой цензурированный снизу ряд. Точнее говоря, значение количества продаж является неотрицательной, а потому, использование нецензурированных эконометрических моделей может привести к тому, что значения оцениваемых параметров при переменных будут смещенными, чтобы при наблюдениях с неблагоприятными условиями для какого-либо неотрицательного значения продаж, оценка объясняющей переменной не слишком сильно отдалялась от нуля в отрицательную часть, тем самым ухудшая метрику, которую выбирает исследователь.

Также, значение количества продаж является дискретным, а вместе с этим, имеет, обращаясь к микроэкономической теории, которая предлагает условную единицу измерения полезности блага, кардиналистский смысл, нежели ординалистский, то есть, 4 продажи дважды больше, чем 2 продажи.

Поэтому класс моделей для так называемых данных подсчета является хорошим выбором для имеющейся выборки. Модели, основанные на распределении Пуассона, удовлетворяют обоим этим свойствам, поэтому, в качестве эконометрических моделей будут применяться именно они.

Обозначим объясняемую переменную через , принимающую целочисленные значения. Наша цель состоит в том, чтобы объяснить распределение или ожидаемое значение , учитывая набор характеристик . Предположим, что матожидание , заданное , задается следующей формулой:

, (2.1)

где – набор неизвестных параметров. Поскольку неотрицателен, мы выбираем функциональную форму, которая производит неотрицательные условные ожидания. Приведенное выше предположение связывает ожидаемый результат с индивидуальными характеристиками в , но не полностью описывает распределение. Если мы хотим определить вероятность данного результата, например, , необходимы дополнительные предположения.

Распространенное предположение в моделях данных подсчета состоит в том, что для заданного переменная подсчета имеет распределение Пуассона с математическим ожиданием Это означает, что условная функция вероятности по , задается формулой

|  |  |
| --- | --- |
|  | (2.2) |

Оценка находится с помощью метода максимального правдоподобия. Если распределение Пуассона соответствует выборке и предполагается, что у нас есть случайная выборка и , это дает последовательную, асимптотически эффективную и асимптотически нормальную оценку для .

Важным уточнением при использовании распределения Пуассона является то, что оно автоматически подразумевает, что условная дисперсия также равна . Это условие называется equidispersion и иллюстрирует ограничительный характер распределения Пуассона. Во многих приложениях равенство условного среднего и дисперсии распределения отклоняется, в том числе в наших данных. Однако, можно последовательно оценить условное среднее значение, не указывая условное распределение. Фактически, модель регрессии Пуассона способна сделать это, даже если распределение Пуассона не ложится на данные. Это связано с тем, что условия первого порядка задачи максимального правдоподобия справедливы в более общем плане, так что мы можем получить состоятельную оценку для , используя метод квазимаксимального правдоподобия (QMLE). Это означает, что мы решаем обычную задачу максимального правдоподобия, но корректируем способ вычисления стандартных ошибок.

Чтобы проиллюстрировать метод квазимаксимального правдоподобия, рассмотрим логарифм функции правдоподобия модели регрессии Пуассона

|  |  |
| --- | --- |
|  | (2.3) |

Каждое слагаемое в сумме по принято называть вкладом правдоподобия. Последний член в этом выражении обычно отбрасывается, поскольку он не зависит от неизвестных параметров.

Условия первого порядка максимизации относительно задаются следующим образом:

|  |  |
| --- | --- |
|  | (2.4) |

Мы можем интерпретировать данное выражение как условия выборочного момента, соответствующие набору условий ортогональности , слабой форме отсутствия эндогенности. Точно также, как ранее мы обозначили элементы суммы (2.3) за вклад правдоподобия, элементы в сумме (2.4) принято в русской эконометрической литературе называть метками. В результате оценка, которая максимизирует , в целом согласуется с условием (2.1), даже если , условно по , не имеет распределения Пуассона. То есть, даже если MLE оценка основана на неправильной функции правдоподобия, но можно утверждать, что она состоятельная на основе справедливости (2.4), то есть равенства суммы меток нулю, то такая оценка называется оценкой QMLE.

Используя информационную матрицу Фишера, которая будет обозначена за , мы можем получить, при выполнении определенных условий регулярности, ковариационную матрицу оценок MLE :

|  |  |
| --- | --- |
|  | (2.5) |

Информационная матрица Фишера отражает кривизну функции логарифма правдоподобия, ясно, что чем выше эта кривизна, тем MLE-оценка точнее. Говоря нестрого, отражает матожидание количества информации о векторе неизвестных параметров, содержащегося в произвольном наблюдении.

Однако, для оценки методом квазимаксимального правдоподобия , подходящей ковариационной матрицей является следующее выражение:

|  |  |
| --- | --- |
|  | (2.6) |

где

|  |  |
| --- | --- |
|  | (2.7) |

Эти ковариационные матрицы можно легко оценить, заменив ожидания средними значениями для выборки, а вектор неизвестных параметров – его оценкой MLE. Ковариационная матрица QMLE аналогична ковариационной матрице Уайта, используемой для модели линейной регрессии. Вместе с этим, можно заметить, что если функция правдоподобия специфицирована правильно, то есть данные действительно подпадают под распределение Пуассона, тогда , и оценка MLE совпадает с QMLE.

В случае, когда , то есть, когда наличествует overdispersion, дисперсия оценки квазимаксимального правдоподобия может быть намного больше, чем предполагается в (2.5). Поэтому полезны альтернативные, более общие модели подсчета.

Одной из альтернатив является применение NegBin I, от полного Negative Binomial I, который является частным случаем отрицательного биномиального распределения. Эта модель подразумевает наличие некоторого параметра , который позволяет overdispersion:

|  |  |
| --- | --- |
|  | (2.8) |

Понятно, что оценки для NegBin I являются состоятельными, только если (2.8) действительно соответствует действительности, и тогда эти оценки более эффективны, чем оценки Пуассона.

Большое обобщение представляет собой NegBin II, которая предполагает:

|  |  |
| --- | --- |
|  | (2.9) |

где параметр показывает, как дисперсия возрастает в зависимости от условного матожидания.

В отличие от модели NegBin I, оценка максимального правдоподобия для модели NegBin II устойчива к ошибкам в распределении. Таким образом, при условии, что условное среднее задано правильно, оценка максимального правдоподобия NegBin II состоятельна.

Вместе с этим, как уже говорилось и демонстрировалось графически ранее, количество наблюдений, где значение продаж по SKU принимает нулевое значение намного больше, чем это допускает распределение Пуассона. Грубо говоря, количество нулевых значений раздуто. Именно для такого случая была придумана следующая модификация, которая получила название Zero-Inflated Poisson (ZIP) для Пуассоновской регрессии и Zero-Inflated Negative Binomial (ZINB) для NegBin II модели.

Интуиция, лежащая в основе модели ZIP, заключается в том, что существует первоначальный лежащий в основе процесс, который определяет, является ли количество продаж нулевым или ненулевым. Как только количество определяется как ненулевое, обычная Пуассоновская регрессия переходит к определению его фактического ненулевого значения. Однако, второй процесс может привести к тому, что определение фактического значения продаж может все равно оказаться нулевым. Первый процесс, генерирующий дополнительные нули – модель бинарного выбора, чаще всего logit регрессия. В итоге ZIP модель состоит из двух следующих частей:

|  |  |
| --- | --- |
|  | (2.10) |
|  | (2.11) |

где – значение вероятности излишних нулей. С математическим ожиданием и дисперсией . Оценкой для выступает следующее значение, оцененное методом моментов:

|  |  |
| --- | --- |
|  | (2.12) |

где – среднее значение выборки, а – дисперсия выборки.

Стоит отметить, что модель ZIP не принадлежит к классу nested моделей, в ней оценка значения продажи в день производится одновременно с обоими условиями.

Последней моделью, которая будет рассмотрена в этом разделе, будет Zero-Inflated Negative Binomial (ZINB). Не вдаваясь в теоретические основания данной модели, лишь уточним, что распределение ZINB модели представляет собой смесь бинарного распределения, которое вырождается при нуле, и обычного распределения подсчета, такого как отрицательное биномиальное распределение.

Опираясь на опытную оценку эконометрического сообщества, наличие большого количества нулей не обязательно означает, что необходимо использования Zero-Inflated модели, прежде всего следует рассмотреть более примитивные модели Poisson regression и NegBin II.

В результате, на имеющихся данных будут оцениваться векторные Poisson regression, NegBin II, ZIP и ZINB модели по всем SKU, сравнительный анализ которых будет проводиться по метрике суммы корней среднеквадратических ошибок по всем SKU. Здесь, для увеличения вычислительной скорости будет использоваться параллельное вычисление имеющихся моделей благодаря их определению матрицы объясняющих переменных для каждого SKU отдельно. Вместе с этим, применяя итерационный отбор факторов путем ‘снизу-вверх’ мы добиваемся последующей оценки только значимых для модели переменных на уровне значимости 1e-4, который был выбран опытным путем в силу того, что приведенные выше алгоритмы довольно быстро переобучаются, запоминая тренировочную выборку.

## 2.2. Алгоритмы машинного обучения, в основании которых лежит алгоритм регрессионного дерева решений

С точки зрения алгоритмов машинного обучения задача предсказания количества продаж в день хорошо ложится на регрессионную задачу дерева решений. Дадим определение решающего дерева. Решающее дерево – рекурсивный алгоритм машинного обучения, целью которого является оптимизация задаваемой метрики с помощью разделения имеющейся выборки на две части в соответствии с выбранным оптимальным условием, предикатом. То есть к каждой внутренней вершине υ приписан предикат: , а к каждой листовой вершине υ приписан прогноз – область значений. Для каждого объекта выборки движение начинается из корня до листа.

Вообще говоря, предикат может иметь любую структуру, если с помощью него можно произвести разбиение, однако на практике используют простое сравнение с пороговым значением t: . Проходя через это узловое место дерева объекты выборки, удовлетворяющие предикату, переходят в правое ветвление , иначе – в левое . Выученная функция получается кусочно-постоянной, а значит, производная равна нулю везде, где задана, поэтому градиентные методы и функциональная оптимизация невозможна. Решающее дерево очень быстро запоминает выборку и переобучается, поэтому очень важно проводить кроссвалидацию и находить оптимальные значения гипер-параметров максимальной глубины дерева, количества минимальных объектов в листе и т.д.

Мы сосредоточимся лишь на регрессионных моделях, построенных на деревьях решений, не рассматривая задачу классификации.

Для задачи регрессии задается функция потерь, например MAE или MSE, которая определяет критерий ветвления – функция, которая оценивает то, насколько уменьшится функция потерь, если результирующие два листа являются конечными, по сравнению с ситуацией, когда исходная вершина сама является листом. Выбирается разделение, которое дает наиболее значительное улучшение. Само же улучшение определяется с позиции информативности (impurity):

|  |  |
| --- | --- |
|  | (2.13) |

Откуда можно найти сам критерий ветвления:

|  |  |
| --- | --- |
|  | (2.14) |

Для задачи минимизации MSE, на которую мы будем опираться, потому как нам нужно, чтобы дерево хорошо научилось обращаться со значениями продаж в период скидок, где условная дисперсия больше, информативность листа может быть определена следующим образом:

|  |  |
| --- | --- |
|  | (2.15) |

Имеющийся алгоритм дерева решений можно значимо улучшить с помощью применения ансамблей. Всего выделяют три основных класса ансамблей, применяемых в машинном обучении: bagging (bootstrap aggregation), boosting и stacking. Ансамбль использует несколько базовых алгоритмов обучения для получения лучшей прогностической производительности, чем можно было бы получить с помощью любого из составляющих базовых алгоритмов обучения в одиночку.

Идея бэггинга заключается в уменьшении разброса при неизменности смещения значений путем суммирования значения предсказаний базовых алгоритмов и деления на количество базовых значений, алгебраически:

|  |  |
| --- | --- |
|  | (2.16) |

где – значение предсказания бэггинга, отвечает за значение предсказания i-го базового алгоритма. При этом наглядно видно, что значение среднего не изменится, так как матожидание суммы всегда равно сумме матожиданий, однако значение дисперсии:

|  |  |
| --- | --- |
|  | (2.17) |

Где, если предположить, что базовые алгоритмы некоррелированы, тогда дисперсия композиции уменьшается в раз. Однако, в реальности добиться последнего условия практически невозможно, ведь базовые алгоритмы учат примерно похожую зависимость на пересекающихся данных. Но, оказывается, что строгое выполнение некоррелированности не обязательно, достаточно определенной степени расхождения самих базовых алгоритмов для того, чтобы данный вид ансамбля удачно работал.

В целом, смысл уменьшения дисперсии состоит в том, что, так как ошибка разложима на чистый шум, смещение и разброс, что доказывается при помощи bias-variance decomposition [Unified Bias-Variance Decomposition… ], уменьшая дисперсию, мы уменьшаем ошибки, улучшая значение функции потерь.

Бэггинг над алгоритмом решающего дерева вместе с использованием метода случайных подпространств называется случайным лесом (random forest).

При кроссвалидации случайного леса важно учитывать такой гипер-параметр как количество базовых алгоритмов, а также все остальные параметры, которые перебираются при кроссвалидации дерева решений. Мы будем находить наиболее оптимальную модель относительно показателя MSE на тестовой выборке для каждого SKU, пробегая следующие гипер-параметры: максимальная глубина - [3, 4, 5, 6, 7, 8, 9, None], количество базовых алгоритмов - [3, 6, 9, 12, 15, 18, 20, 30, 40, 50], количество факторов в базовом алгоритме - ['auto', 1 / 3]. Гипер-параметры, которые не подвергаются кроссвалидации: минимальное количество выборок, необходимых для нахождения в конечной вершине – 2, минимальное количество выборок, необходимое для разделения внутренней вершины – 4.

Для увеличения скорости вычисления можно использовать параллельное программирование, которое становится возможным, потому что базовые алгоритмы могут быть оценены независимо.

Второй названый выше класс ансамбля – бустинг, тоже получил популярность при работе с алгоритмом дерева решений. Суть заключается в последовательном улучшении значения прогноза путем объяснения разницы между спрогнозированным и действительным значениями.

Имея некоторую функцию потерь, в нашем случае MSE, мы используем композицию из базовых алгоритмов , в нашем случае – деревьев решений:

|  |  |
| --- | --- |
|  | (2.18) |

где

|  |  |
| --- | --- |
|  | (2.19) |
|  | (2.20) |

где – темп обучения (learning rate), принадлежащий интервалу (0, 1), - MSE функция потерь.

Учитывая, что

|  |  |
| --- | --- |
|  | (2.21) |

и, принимая во внимание (2.19), получается, что для всякого очередной алгоритм в ансамбле обучается предсказывать антиградиент функции потерь по прогнозу модели в точке .

Градиентный бустинг над решающими деревьями кратко обозначается за GBDT и является одним из двух самых распространенных алгоритмов машинного обучения, в особенности при работе с неоднородными данными, второй – нейросети.

Применение градиентного бустинга к моделям, имеющим четкую функциональную форму, не распространено. Когда, например, градиентный бустинг выполняется вместе с линейной регрессией, это не что иное, как еще одна линейная модель поверх существующей линейной модели. Интуитивно это можно понять как добавление чего-то к уже найденным коэффициентам, и, если линейная регрессия уже нашла наилучшие коэффициенты, это будет бесполезно в силу того, что .

Кроссвалидация GBDT, помимо подбора гипер-параметров деревья решений, сосредотачивается на подборе подходящего гипер-параметра темпа обучения. То есть, в соответствии с нашими данными и поставленной задачей, гипер-параметры кроссвалидации для GBDT моделей по всем SKU содержат следующее: темп обучения - [0.005, 0.01, 0.03, 0.06, 0.1, 0.15, 0.3], глубина дерева - [None, 3, 4, 5, 6, 7], максимальное количество итераций – 500. Данный процесс построения GBDT модели завершается, как только следующая рассчитанная итерация не увеличивает значение выбранной метрики.

## 2.3. Формирование алгоритма долгосрочного прогнозирования

Для осуществления поставленной цели в лице калькуляции долгосрочного прогноза, определим сам алгоритм предлагаемого прогнозирования.

Так как под прогнозированием понимается формирование оценок количества продаваемых единиц SKU при отсутствии информации о реальных значения предыдущих продаж, когда номер наблюдения включает в себя только лаги, которые выходят за рамки границ обучаемой выборки, то прогнозирование количества продаж по вектору таргетируемых SKU должен определяться условно по имеющейся предыстории .

И, как уже говорилось ранее, среди всех объясняющих переменных при осуществлении прогноза мы не имеем знание о реальном неценовом спросе на все SKU, а также на реальные значения статистик неценового количества продаж за последнюю неделю по всем SKU.

Размер тестовой выборки, на которой будет осуществляться прогнозирование составляет 15% общих данных. Имея общее количество обозреваемых наблюдений в величину 2173, индекс начала прогноза – 1847, который мы будем обозначать за . Временной индекс, который будет пробегать по наблюдениям тестовой части выборки, обозначим за .

Для того, чтобы иметь оцененный прогнозом лаг значения неценового спроса на какой-либо SKU, нам необходимо взять лаг оцененного моделью значения количества продаж по данному SKU и отнять оценку объясненного ценой этого значения, которая формируется вспомогательной моделью по цене и факта поставок. Вспомогательная модель является той же теоретической моделью, что и оцениваемая модель. То есть, в итоге:

|  |  |
| --- | --- |
|  | (2.22) |
|  | (2.23) |

В свою очередь, имея значения неценового спроса на все SKU, мы можем произвести калькуляцию недельных статистик:

|  |  |
| --- | --- |
|  | (2.24) |
|  | (2.25) |

Где для любого id SKU, включая и id SKU таргетируемой переменной выполняется:

|  |  |
| --- | --- |
|  | (2.26) |

, где определяется вспомогательной моделью по цене и наличию продаж, являясь ее оценкой.

Вычислительный процесс можно ускорить в несколько раз с помощью применения параллельных синхронных процессов, когда на каждом наблюдении изменение матриц объясняющих переменных, которое производится за счет вычисления приводимых в формулах (2.22) – (2.26), а также вычисление самого прогноза для SKU при помощи соответствующей модели можно подвергнуть параллелизации по всем SKU.

# ГЛАВА 3. ОПИСАНИЕ И ОЦЕНКА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

## 3.1. Сравнение результатов рассмотренных моделей на имеющихся данных

Оценивание приводимых выше эконометрических моделей, основанных на распределении Пуассона: Poisson regression, Zero-Inflated Poisson, NegBin II, а также алгоритмов машинного обучения, в основании которых лежит алгоритм регрессионного дерева решений: Random Forest, Gradient Boosting Decision Trees, которое проходило в вектором виде, базируясь на созданном алгоритме, завершалось вычислением ключевой, обозначенной ранее, метрики – mean squared error (MSE) – среднеквадратичная разница между расчетными значениями и фактическим значением, по тренировочной выборке и тестовой части по каждому SKU в отдельности.

Далее, вычисляется суммарное MSE по тестовой выборке по всем SKU с помощью простой алгебраической суммы рассчитанных MSE. Впоследствии, применяя алгоритм векторного прогнозирования, который был сформулирован выше, также вычисляем значение MSE, которое обозначим за MSE forecast, и, как и в случае с тестовой выборкой, рассчитываем суммарное значение метрики.

По причине наличия достаточного малого количества наблюдений и сложности самой модели, Zero-Inflated NegBin II модель не удалось оценить хоть как-то состоятельно, из-за чего результаты данной модели не будут отображены в данной главе.

Представляются результаты по первому классу моделей (таб. 1).

Таблица 1 – Результаты оценки моделей Poisson regression, Zero-Inflated Poisson, NegBin II

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID SKU | Vector Poisson | | Vector ZIP | | Vector NegBin II | |
| MSE test | MSE forecast | MSE test | MSE forecast | MSE test | MSE forecast |
| 46135 | 8,608 | 10,781 | 8,007 | 11,678 | 10,122 | 12,545 |
| 46902 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 59042 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 75320 | 3,935 | 5,218 | 3,924 | 5,319 | 3,563 | 4,989 |
| 201676 | 21,961 | 25,727 | 21,368 | 25,598 | 22,667 | 27,752 |
| 362058 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 415514 | 7,511 | 10,043 | 7,167 | 10,014 | 8,299 | 11,322 |
| 415824 | 1,127 | 1,522 | 1,130 | 1,521 | 0,806 | 1,342 |
| 419020 | 11,146 | 13,828 | 11,086 | 13,814 | 9,955 | 14,960 |
| 437966 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 437982 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 459894 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 459896 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 461504 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 468496 | 2,050 | 2,216 | 2,027 | 2,200 | 2,046 | 2,244 |
| 468558 | 7,639 | 9,463 | 7,699 | 9,439 | 7,987 | 10,881 |
| 471503 | 9,170 | 11,871 | 9,840 | 11,999 | 9,267 | 12,738 |
| 481677 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 489135 | 30,269 | 32,445 | 29,742 | 32,391 | 31,240 | 35,151 |
| 600761 | 2,327 | 3,228 | 2,051 | 3,092 | 2,147 | 3,236 |
| Total MSE | 105,744 | 126,342 | 104,041 | 127,065 | 108,099 | 137,160 |
| Total RMSE | 32,441 | 35,714 | 31,498 | 35,136 | 31,852 | 36,286 |
| Total MSE train | 62,937 | | 61,466 | | 136,681 | |

А также по моделям машинного обучения (таб. 2).

Таблица 2 – Результаты оценки моделей Random Forest, GBDT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID SKU | Vector Random Forest | | Vector GBDT | |
| MSE test | MSE forecast | MSE test | MSE forecast |
| 46135 | 24,945 | 26,555 | 3,470 | 3,467 |
| 46902 | 0,000 | 0,000 | 0,000 | 0,000 |
| 59042 | 0,000 | 0,000 | 0,000 | 0,000 |
| 75320 | 4,091 | 4,241 | 4,957 | 5,044 |
| 201676 | 18,760 | 21,074 | 19,605 | 20,056 |
| 362058 | 0,000 | 0,000 | 0,000 | 0,000 |
| 415514 | 3,592 | 4,176 | 4,890 | 5,696 |
| 415824 | 0,946 | 3,595 | 1,051 | 1,134 |
| 419020 | 18,240 | 37,355 | 9,099 | 9,142 |
| 437966 | 0,000 | 0,000 | 0,000 | 0,000 |
| 437982 | 0,000 | 0,000 | 0,000 | 0,000 |
| 459894 | 0,000 | 0,000 | 0,000 | 0,000 |
| 459896 | 0,000 | 0,000 | 0,000 | 0,000 |
| 461504 | 0,000 | 0,000 | 0,000 | 0,000 |
| 468496 | 2,823 | 2,997 | 2,152 | 2,204 |
| 468558 | 13,493 | 19,853 | 7,251 | 8,406 |
| 471503 | 24,868 | 25,522 | 18,833 | 18,426 |
| 481677 | 0,000 | 0,000 | 0,000 | 0,000 |
| 489135 | 10,903 | 13,764 | 11,417 | 23,198 |
| 600761 | 1,352 | 1,440 | 1,726 | 1,911 |
| Total MSE | 124,013 | 160,572 | 84,452 | 98,684 |
| Total RMSE | 33,933 | 38,804 | 30,505 | 34,579 |
| Total MSE train | 48,082 | | 44,961 | |

Видно, что имеются SKU, значение MSE по которым равно 0, например, 46902, 59042, которые соответствуют 2 и 3 номеру. Причина тому в том, что на протяжении всего периода тестовой части данные SKU не имели поставок, а значит, значения продаж по ним были соответственно равны нулю, что было учтено моделями, в которых находилась переменная наличия поставок. Также, например, можно понять, что SKU 600761, 415824 поставлялись только определенную малую часть периода.

Как можно заметить, рассматривая значения MSE test моделей, наиболее удачной моделью оказался GBDT, после которого идут, в соответствии с порядком возрастания MSE test, ZIP, Poisson Regression, NegBin II, и, наконец, Random Forest, результат которого на 47% хуже, чем у лучшей рассмотренной модели, хотя обе модели имеют базовым алгоритмом дерево решений.

Вместе с этим, можно заметить, что модели, основанные на Пуассоновском распределении, имеют очень схожие результаты, хотя NegBin II делает прогноз значимо хуже остальных. Из чего можно сделать вывод о том, что даже наличие чрезмерной выборочной дисперсии относительно выборочного среднего не гарантирует более лучших по сравнению с Poisson regression результатов. Также, ZIP модель не показала какого-либо преимущества при оценке, что свидетельствует о том, проводя аналогию с выводом выше, что большое количество нулей в объясняющей переменной не гарантирует более лучших по сравнению с Poisson regression результатов.

Вдобавок, рассмотрим на конкретном SKU графическое изображение кривой фактического количества продаж и кривых оцененных значений моделей, основанных на распределении Пуассона (рис. 3), а также рассмотренных моделей машинного обучения (рис. 4).

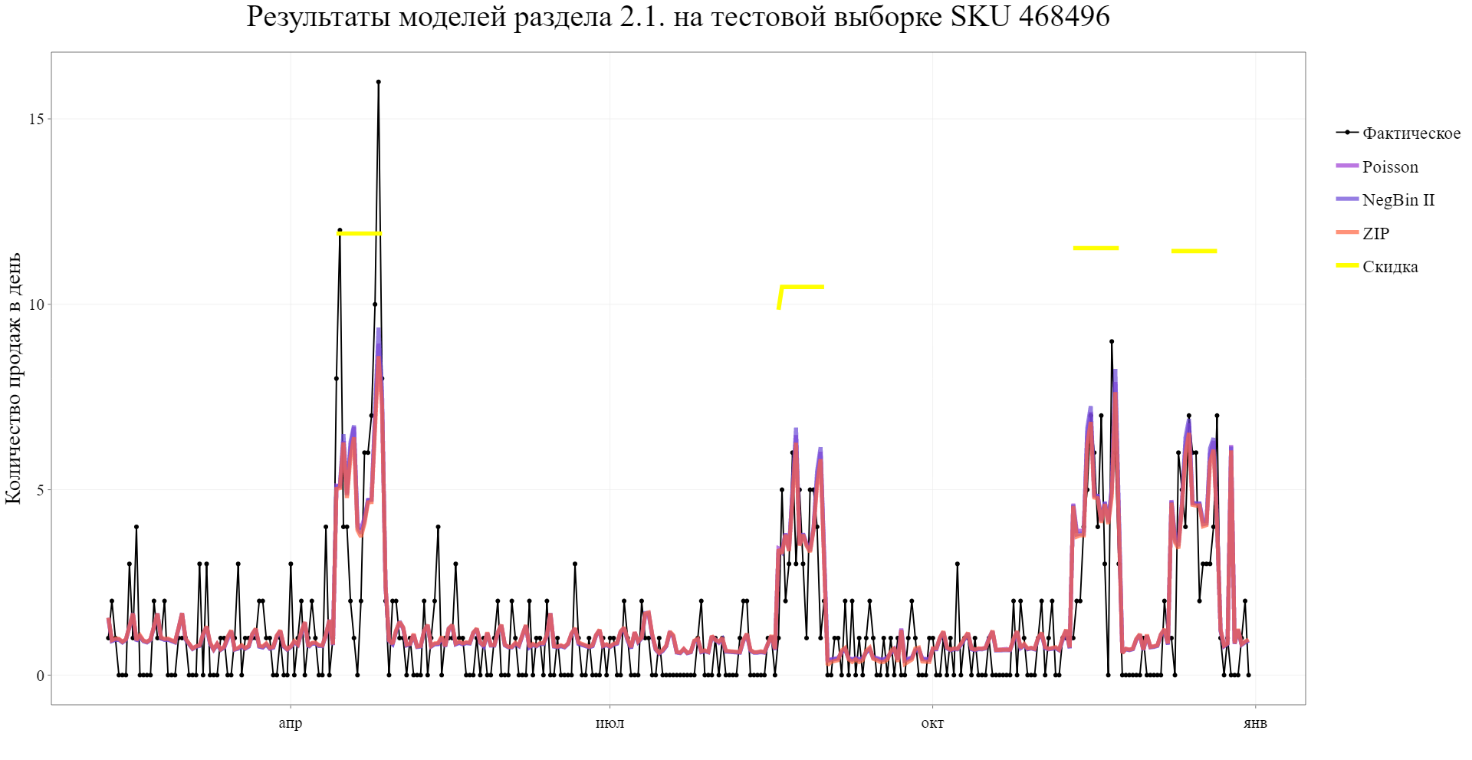


Рисунок 3 – Результаты моделей, основанных на распределении Пуассона, на тестовой выборке SKU 467496

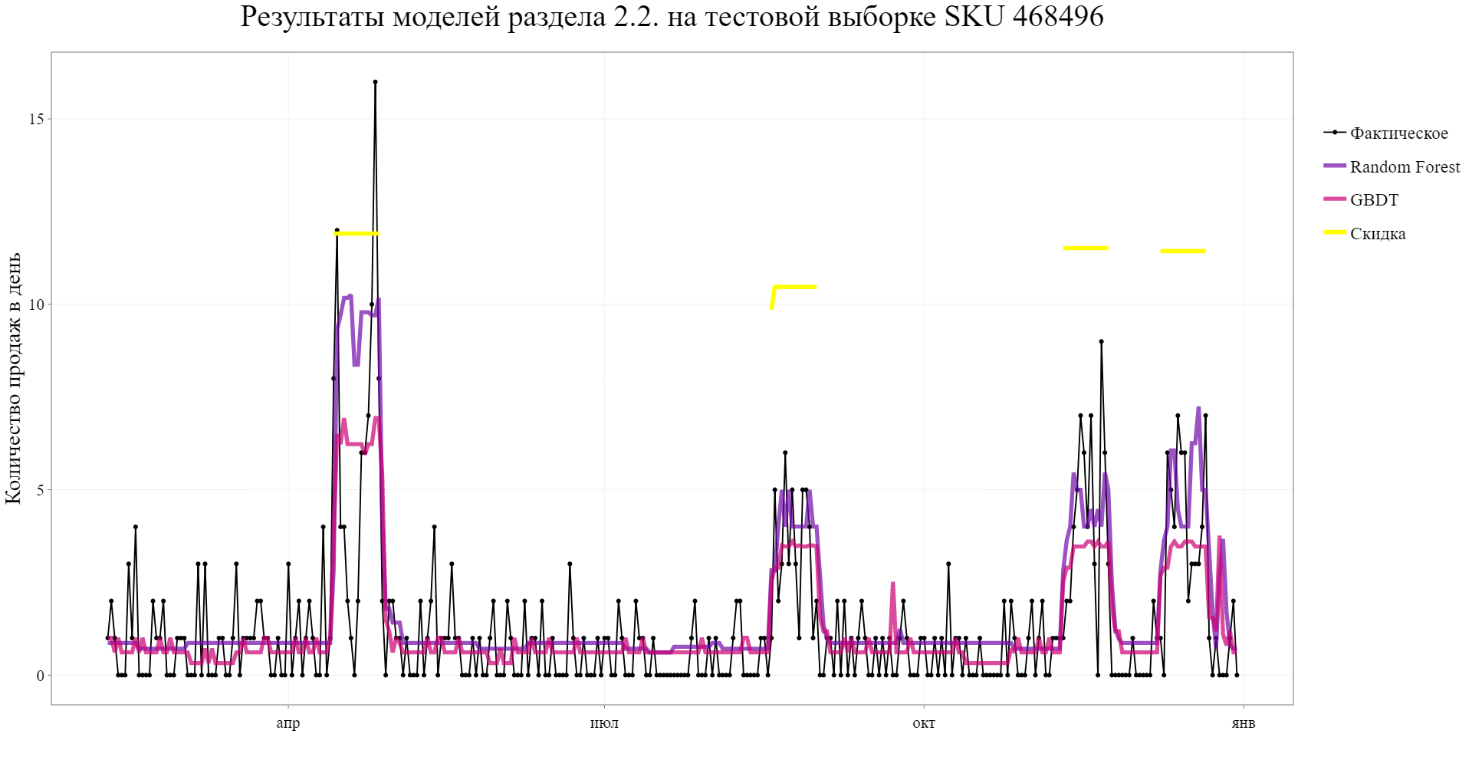


Рисунок 4 – Результаты моделей Random Forest и Gradient Boosting Decision Trees на тестовой выборке SKU 467496

Как можно увидеть, результаты моделей, основанных на распределении Пуассона практически, не различаются, разве что NegBin II имеет чуть больший разброс, что является прямым следствием предположений модели. Poisson regression сама оказалась достаточно оптимальна для имеющихся данных, потому дальнейшие модификации не оказали значимого влияния на результаты.

Однако, результаты RF и GBDT моделей довольно разные, как можно заметить, из-за своего построения, GBDT в дни с обычным значением цены более вариабельна, чем RF, который предсказывает некоторые вариации среднего значения продаж в день, проводя ломаную линию. Вместе с этим, RF предсказывает для данного SKU более повышенный спрос в дни скидок, чем GBDT, однако, показывает более высокую вариацию, чем последняя модель.

Проведем небольшой анализ результатов Poisson regression по SKU 467496 (таб. 3).

Таблица 3 – Результаты оценки Poisson regression по SKU 467496

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SKU 467496 – Poisson Regression Results – Pseudo R-squ.: 0.6026 | | | | | | |
| variable | coef | std error | z | P > |z| | [0.025 | 0.975] |
| tar\_price | -0,0718 | 0,002 | -37,944 | 0 | -0,075 | -0,068 |
| tar\_non\_price\_1 | 0,0363 | 0,006 | 5,637 | 0 | 0,024 | 0,049 |
| is\_supplied\_471503 | 0,3225 | 0,149 | 2,166 | 0,03 | 0,031 | 0,614 |
| is\_supplied\_481677 | 0,4061 | 0,055 | 7,339 | 0 | 0,298 | 0,515 |
| weekday\_Friday | 0,3260 | 0,065 | 5,051 | 0 | 0,199 | 0,452 |
| weekday\_Saturday | 0,4551 | 0,06 | 7,553 | 0 | 0,337 | 0,573 |
| month\_September | -0,3898 | 0,085 | -4,588 | 0 | -0,556 | -0,223 |
| non\_price\_demand\_481677\_lag | 0,0190 | 0,011 | 1,771 | 0,077 | -0,002 | 0,04 |
| non\_price\_demand\_600761\_lag | 0,0360 | 0,008 | 4,267 | 0 | 0,019 | 0,053 |
| tar\_is\_supp | 4,7338 | 0,241 | 19,607 | 0 | 4,261 | 5,207 |

В отличие от модели линейной регрессии, модели бинарного выбора не имеют единой меры качества подгонки данных под модель, но существуют разные меры оценки степени объяснения. Часто показатели качества подгонки данных с помощью модели неявно или явно основаны на сравнении с моделью, которая содержит только константу в качестве объясняющей переменной. Чем больше разница между двумя значениями логарифма правдоподобия, тем лучше расширенная модель дополняет очень ограниченную модель. Одной из таких мер оценки является скорректированный псевдо-R2, значение которого приведено в таблице (таб. 3). Для переменных, которые являются непрерывными, интерпретация значений параметров проводится в соответствии с приводимой ниже формулой.

|  |  |
| --- | --- |
|  | (3.1) |

Как видно, данная формулировка напрямую дает значение эластичности зависимой переменной по объясняющей. То есть, при прочих равных, он измеряет относительное изменение ожидаемого значения количества продаж в день при изменении объясняющей переменной на 1%. Так, прямая ценовая эластичность для SKU 467496 составляет -0,0718.

Для параметров бинарных переменных интерпретация параметров вытекает из следующей формулы:

|  |  |
| --- | --- |
|  | (3.1) |

Для значений малых значений параметров справедливо , которое можно применить к переменным наличия поставок остальных SKU, а также к dummy-переменным дня и месяца. Например, значение параметра при переменной субботы указывает на то, что ожидаемое значение количества продаж увеличивается примерно на 45%, если переменная индикатора изменяется с 0 на 1.

Также, рассмотрим визуализацию прогноза по данному SKU (рис. 5) по лучшей из всех рассматриваемых моделей – GBDT.

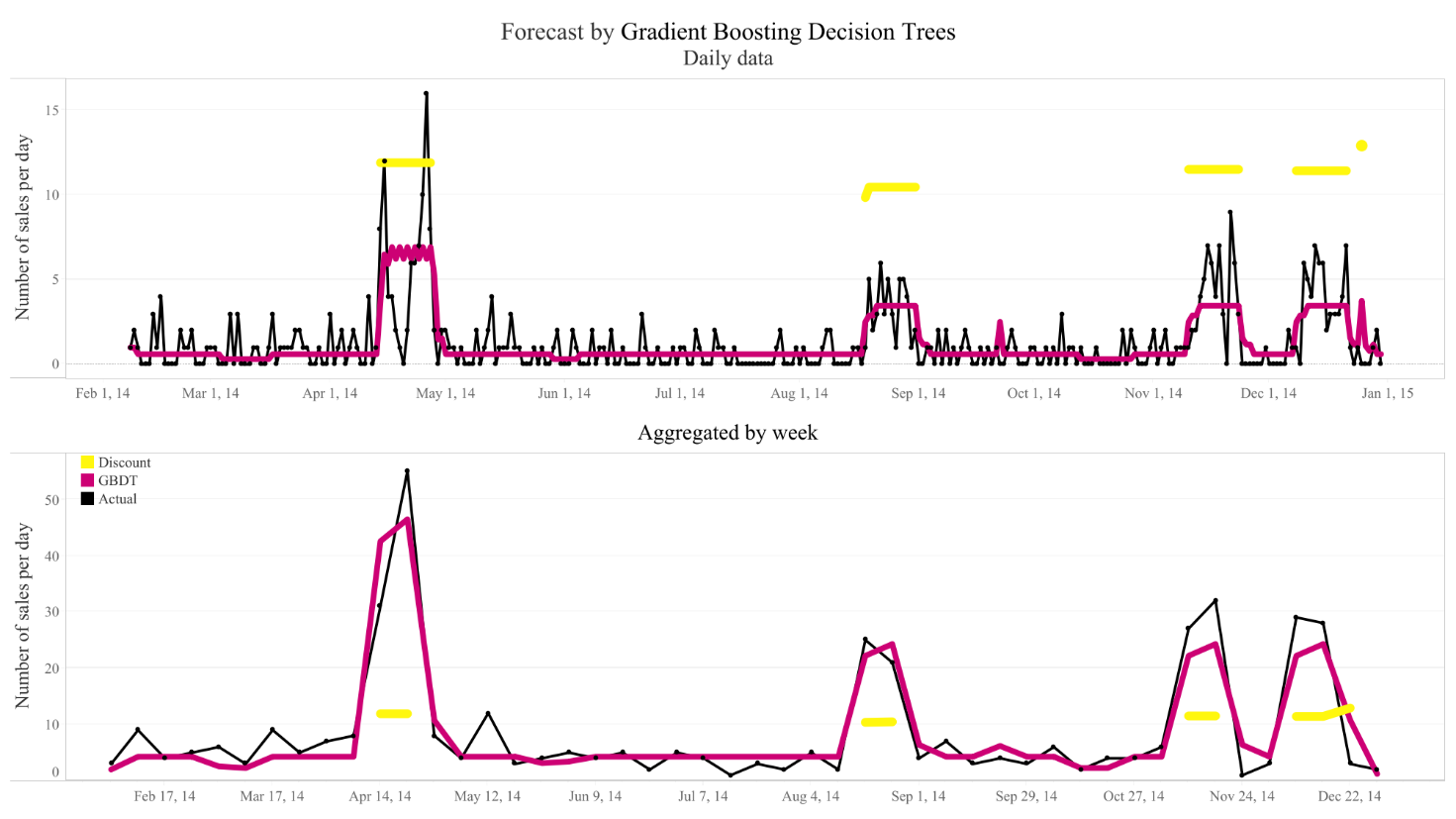


Рисунок 5 – Результаты прогнозирования по GBDT на тестовой выборке SKU 467496 в дневном и агрегированном понедельном формате

Как видно, результаты прогнозирования являются достаточно точными, а сам алгоритм не сильно полагается на значения предыдущих продаж, из-за чего накапливаемая ошибка отсутствует. Агрегированные понедельно результаты представляются еще более точными, потому как эффект случайного спроса, который практически невозможно отследить, нивелируется.

Если предположить, что текущие результаты могут быть масштабированы на другие товарные категории при таком же относительном уровне ошибки, то розничный торговец может прогнозировать свои денежные потоки и формировать свою бюджет на год с большим уровнем точности.

## 3.2. Несовершенство представленных моделей и неэффективность потенциальной оптимизации по ценам

Одна из проблем, которая встает на пути желаемой задачи оптимизации по ценам, заключается в том, что обучающиеся алгоритмы не имеют представления о имеющихся ценах на товарную категорию в других близлежащих магазинах. Выражается же эта проблема в том, что при оптимизации цен, имея обученное представление о взаимосвязях цены и продаваемого количества, модель может повысить условно стабильные цены на SKU товарной категории с целью максимизации выручки, не зная о том, что цены в близлежащих магазинах действительно меньше и спрос на товарную категорию самого магазина может значительно упасть.

Однако, данная проблема может быть не столь значительна, потому как перекрестная эластичность на некоторые товары не входящих в перечень продовольственных товаров первой необходимости может оказаться достаточно малой. С другой стороны, мы можем избежать проблему повышения цен простым ограничением на не отрицательность значения скидки.

Другой проблемой может являться то, что модели не знают о том, к какому изменению количества продаж может привести долгое наличие скидки, например, в период больше месяца. Связано это с тем, что данные, на которых обучается модель, сами по себе не являются случайными, а содержат продуманный план по продаже соков, который бы учитывал то, что долгий период наличия скидки, при условии относительно локальной ограниченности спроса, может негативно влиять на количество продаж, что можно связать с психологическим эффектом скидки – ощущением срочности. По причине того, что мы не имеем действительно долгого периода скидок, как минимум для большинства SKU, модель не может уловить этот эффект, который бы, например, мог выражаться через значимый отрицательный коэффициент для количества продаж за месяц по таргетируемому SKU, который принимает большие значения в долгий период скидок.

В основном, другие всевозможный аргументы в несостоятельной потенциальной оптимизации по ценам состоят в том, что модель не учитывает психологические аспекты скидки. Так, например, ощущение срочности, про которое мы говорили, заставляет розничные сети прибегать к созданию окон без скидок. Привычка к скидкам тоже может негативно сказываться на объем продаж - если их ожидания не оправдаются, клиенты идут покупать у конкурентов. Отсутствие ожидаемой скидки может разочаровывать покупателя, поэтому рекомендуется варьировать виды скидок. Таким образом, можно будет сохранить поток покупателей. Однако, представленные модели не могут учесть этот факт, а потому, скорее всего, ценовая оптимизация приведет к тому, что полученные цены будут являть собой вектор перманентных значений псевдо-оптимальной скидки для SKU для всего прогнозируемого периода.

# ЗАКЛЮЧЕНИЕ

Современные методы математического моделирования, помогают выявить закономерности в данных, которые тяжело интуитивно отследить и тем более, крайне тяжело эффективно измерить.

Имея начальные данные, которые содержат в себе только дневные значения продаж и цен по товарным единицам, можно, построив качественную базу данных, использовать сложно структурированные прикладные модели, с помощью которых можно с достаточной аккуратностью составлять прогноз по количеству продаж на определенный долгосрочный период времени, и, которые могут стать большим подспорьем в формировании прогнозов по денежным потокам по операционной деятельности и формировании бюджета.

Однако, представленные в настоящей курсовой работе модели и алгоритмы все равно далеки от эталонных. Они не учитывают многие факторы, которые коренным образом влияют на количество продаж, а потенциально возможная оптимизация выручки скорее всего приведет к несостоятельным результатам.

Решением же возникающих проблем может быть, во-первых, выбор более содержательного изначального набора данных, с помощью которого предполагаемые модели могли бы научиться более вариативным связям, например, понимать и улавливать эффект срочности, во-вторых, создание более информативной матрицы объясняющих переменных, а также, применение более гибких, передовых моделей машинного обучения.

# СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

Андреас К., Мюллер Д. К. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными: моногр. / Мюллер Андреас. - М.: Альфа-книга, 2017. – 697 c.

Ванде Д. Python для сложных задач. Наука о данных и машинное обучение. Руководство / Плас Джейк Вандер. - М.: Питер, 2018. - 527 c.

Вербик М. Путеводитель по современной эконометрике // М.: Научная книга, 2008. - 616 с.

Елистрова Е., Губко П. Ансамбли в машинном обучении // Учебник по машинному обучению от школы анализа данных [электронный источник]. - URL: https://ml-handbook.ru/chapters/ensembles/intro (дата обращения: 21.04.2022)

Индекс потребительских цен на фруктовые соки, Пермский край, в % к предыдущему месяцу // Центральная база статистических данных, Федеральная служба государственной статистики [электронный источник]. - URL: https://www.gks.ru/dbscripts/cbsd\_internal/DBInet.cgi?pl=1902001 (дата обращения: 03.05.2021)

Лунев К., Елистрова Е. Градиентный бустинг // Учебник по машинному обучению от школы анализа данных [электронный источник]. - URL: https://ml-handbook.ru/chapters/grad\_boost/intro (дата обращения: 19.05.2022)

Синицин Ф. Решающие деревья // Учебник по машинному обучению от школы анализа данных [электронный источник]. URL: https://ml-handbook.ru/chapters/decision\_tree/intro (дата обращения: 014.03.2022)

Соколов Е. А. Лекция 8 – Бэггинг, случайные леса и разложение ошибки на смещение и разброс // НИУ ВШЭ, Факультет компьютерных наук, 2020. – 10 с.

Breiman, Leo. Random Forests. // Machine Learning, 2001. – 5-32 с.

Cameron A. C., Trivedi P. K. Regression Analysis of Count Data, 2nd edition, Econometric Society Monograph No.53, Cambridge University Press, 1998. – 556 c.

Cameron A. C., Windmeijer F. R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization, 1996. – 209-220 c.

CatBoost // Продвинутая библиотека градиентного бустинга на деревьях решений с открытым исходным кодом. [электронный источник] URL: https://catboost.ai/ (дата обращения: 022.03.2022)

Domingos, Pedro. A Unified Bias-Variance Decomposition and its Applications. // In Proc. 17th International Conf. on Machine Learning, 2000.

Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine, 2001. – 39 c.

Hastie T., Tibshirani R., Friedman, J. The Elements of Statistical Learning. // Springer, New York, 2001. – 83-84 c.

James H. Stock, Mark W. Watson - Introduction to Econometrics-Pearson, 2015. – 839 c.

Lambert D. “Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing,” Technometrics, 1992. – 1-14 с.

Sachin D. An Illustrated Guide to the Zero Inflated Poisson Regression Model [электронный источник] – URL: https://towardsdatascience.com/an-illustrated-guide-to-the-zero-inflated-poisson-model-b22833343057 (дата обращения: 05.05.2022)

Veerbek - A Guide to Modern Econometrics, 2000. – 448 c.