**1. Введение**

**1.1 Описание предметной области.**

Прогнозирование, как таковое, появилось очень давно. Ещё с самых первобытных времён людям было интересно, будет ли завтра дождь или пойдёт ли завтра снег. Конечно, первые люди не имели никакого представления о том, как устроена погода и прочее. Поэтому появились боги, благодаря которым, как думали люди, всё на свете и случается. Так зародились мифология, философия и другие, которые могли объяснить и подсказать, что же нужно делать дальше.

Время шло и люди заметили, что окружающий их мир поддаётся определённым законам и последовательностям. Например, что солнце встаёт каждое утро с одной и той же стороны, а садится с противоположной. Что происходит это всегда примерно в одно и то же время. Что привело к тому, что люди стали ещё больше изучать окружающий мир и пытаться понять его.

Когда появились первые зачатки экономики и основы торговли, стало понятно, что, и количество продаваемого продукта, и количество произведённого продукта, можно как-то регулировать.

В какой-то момент люди осознали, что не имеет смысла производить 100 пар обуви, когда в поселении живёт всего 15 человек, ибо это не имеет смысла. 15 пар кто-то купит, а остальные либо останутся стоять и ждать своего покупателя, либо же кто-то проезжий когда-нибудь купит их, что очень на вряд ли. Но это в самом простом примере, если рассматривать данный вопрос в огромных масштабах, можно понять насколько он становится актуальным. Если одна произведённая деталь стоит 5 условных единиц, а продаётся она за 7 условных единиц, но покупают её редко, появляется вопрос, а стоит ли эту делать производить и насколько много на ней можно заработать, а не наоборот, потерять.

Поэтому, управление доходами, всегда было актуальной и, с большой вероятностью, останется одной из самых актуальных прикладных областей исследования операций.

Сейчас, управление доходами – это набор стратегий и подходов, позволяющий распределить имеющиеся ресурсы так, чтобы наилучшим образом удовлетворить спрос на них и получить максимальную прибыль. Благодаря этому, прогноз спроса является очень важной составляющей многих моделей принятия решений, которые используются в области управления доходами.

В сфере услуг оптимальное управление ресурсами можно достичь с помощью контроля наличия в продаже различных продуктов. То есть, если знать, где нужно добавить товара, а где снизить, можно добиться наилучшего получения прибыли. И любая стратегия продаж зависит от спроса на товар. Очевидно, если неточно спрогнозировать продажи, это может привести к огромным убыткам.

И именно по этой причине, так востребованы подходы к моделированию спроса, не ограниченного пределами бронирования и вместимостью. Такой спрос имеет название «общий». Сложность прогнозирования такой спроса в том, что исторические данные – это уровень продаж, а не сам спрос. То есть, возможно, желание купить было у 100 человек, а деталей было всего 30. Остальные 70 человек остались без деталей. То есть спрос больше, чем сами продажи. А это отобразить на каких-то данных невозможно, тк люди, узнав, что деталей нет, обращаться не будут. Таким образом, исторические данные о продажах будут соответствовать цензурированному спросу. Процедура оценивания параметров распределения такого спроса по исходной выборке имеется название – восстановление общего неограниченного спроса, а сами методы для него имеют название – методы восстановления. Довольно трудно измерить общий неограниченный спрос, но при этом, некоторые исследователи утверждают, что использование этих методов улучшает прогноз и увеличивает прибыль. Так, было обнаружено разными учёными, что влияние восстановления спроса на доход может быть от 2 процентов до 12.

Дипломная работа посвящена теме исследования применимости метода бутстрап и других непараметрических методов восстановления распределения спроса на товары. В качестве примера взяты данные магазинов с продажами.

**1.2 Состояние изученности темы**

Существует очень много научных трудов, которые поднимают вопрос восстановления общего неограниченного спроса, оценке и прогнозированию. В частности, в работах идёт обсуждение новых методов восстановления, сравнение точности полученных оценок и анализ влияния выбранных подходов на доход. Но в этих работах мало уделяется внимание сравнению статических свойств, полученных оценок, чувствительность методов к разнообразным ситуациям. Так же существуют методы, которые не воспринимают неполноту информации о цензурировании.

Благодаря этому всему, данные вопросы являются недостаточно раскрытыми и актуальными. Это и послужило сделать выбор такой темы дипломной работы.

**1.3 Формулировка цели и задач исследования**

Целью дипломной работой является - исследовать применимость метода бутстрап и других непараметрических методов для восстановления распределений спроса на товары.

1. Проведён анализ недостатков, преимуществ и особенностей, в частности, метода бутстрап, и других непараметрических методов восстановления распределений спроса
2. Изучение асимптотических свойств метода бутстрап
3. Предложен и обоснован подход восстановления спроса с помощью метода бутстрап.
4. Проведены расчёты, с помощью которых можно сравнить точность оценок, полученных различными методами восстановления спроса.

**1.4 Предмет исследования**

Предметом исследования является восстановление спроса на товар в магазинах по цензурированным данным в рамках управления доходами.

**1.5 Объект исследования**

Объектом исследования являются 54 магазина сети Favorita, которые расположены в Эквадоре.

**2. Управление доходами**

**2.1 Управление доходами. Основные понятия**

В последнее время, такой понятие как «управление доходами» стало не просто словосочетанием, а является прикладной областью исследования операций. Во всём мире, от малых фирм до огромных, происходит повседневных подсчёт прибыли, подсчёт произведённого и проданного товара, и прочего. В связи с техническим процессом это стало делать очень удобно, есть где и как хранить информацию. И появились средства для того, чтобы что-то с этой информацией делать. Появились целые группы аналитиков, маркетологом, менеджеров и прочих, которые хотят увеличить прибыл свою и прибыль компании, в которой они работают. Возникает вопрос, как же это сделать. Естественно, можно спрогнозировать то или иное поведение на рынке и заранее знать, что лучше и как лучше производить, продавать и покупать. Но это же будущее, которое просто так взять и спрогнозировать нельзя, естественно будет какая-то ошибка, но, чем точнее будет прогноз, тем наибольшую прибыль получат все, кто в этом участвовал. Поэтому и начали придумать различные методы прогнозирования спроса, стали их улучшать, исследовать и рассматривать, чтобы добиться наименьшей ошибки.

Существует несколько определений управления доходами. Например, это применение особых к подходов, которые предсказывают действия потребителей в какой-то из сфер и тем самым улучшить цены и тем самым максимизировать свою прибыль. Или же, управление доходами, это – умение применить знания и стратегии для того, чтобы правильно продать правильный товар, правильному человеку, в правильном месте и в правильное время.

Управление доходами включает в себя:

1. Ценовые решения. Как именно выставлять цены, как ими варьировать и прочее.
2. Структурные решения. Под каким форматом продаж выставлять продукты, будь то скидки или наоборот, завышение цен.
3. Количественные решения. Продавать или запрещать продажу. Распределение имеющихся ресурсов между различными сегментами.

Что из этого является наиболее важным и на что нужно акцентировать внимание зависит от сферы деятельности, от конкретной ситуации на рынке и от многих других факторов.

Имеется общая схема работы системы управления доходами, она заключается в том, что существует бесконечный цикл, который состоит из этапов, повторение которого происходит через определённое количество времени. Как часто этот происходит, зависит от многих факторов, таких как: скорость изменения условий бизнеса, объём данных, методы используемые для оптимизации и прогнозирования, значимость решений и другие. Из этого всего можно выделить основные шаги цикла работы системы управления доходами:

1. Во-первых, нужны данные, от которых нужно отталкиваться. Поэтому первым пунктом является сбор данных. Данные могут быть разными, это и исторические данные о спросе, причинных факторах и даже цены. Так как этот пункт очень важен и без него нельзя осуществить последующие пункты, нужно с особой серьёзностью относиться к источнику данных и наилучшим образом им воспользоваться.
2. Во-вторых, это оценка и прогнозирование. Оценка распределения спроса, оценка распределения спроса с использованием методов восстановления, прогнозирование спроса с полученными параметрами. Так же прогнозирование иных не менее важных показателей, например, возвраты, неявки и прочее.
3. В-третьих, все те данные, что получены, нужно оптимизировать. Поэтому, третьим этапом является оптимизация. Оптимизация – определение наилучших значений для параметров регулирования и контроля. То есть, рассчитываются оптимальные пределы создания продукта с учётом выбранной стратегии продаж, выбраны подходящие цены на продукцию, уровень скидок и наценок.
4. В-четвертых, последний шагом является контроль. То есть, контролирование продажность продукции с использованием тех знаний, что были получены на предыдущих шагах.

Таким образом получается, что общая схема работы системы управления доходами выглядит так: собранные данные приходят в блок прогнозирования и оценивания, полученные прогнозируемые значения становятся входными данными для определения оптимальных параметров, далее оптимальные параметры регулирования используются для контроля и обработки запросов. Затем, полученные данные отправляются в историческую базу. После чего цикл замыкается и начинается сначала. Эта процедура указана на Рисунке 1.

# 

Рис. 1. Основные этапы работы системы управления доходами

**2.2 Оценка и прогнозирование спроса.**

Оценка и прогнозирование является обширной темой, которая охватывает многие области научной деятельности, такие как статистику и экономику, технические и компьютерные науки. Некоторые методы основаны на точных статических и математических утверждениях, некоторые эвристические.

Так как под оценкой зачастую подразумевается нахождение параметров модели, которые наилучшим образом описываются заданный наблюдаемый набор данных, а прогнозированием является предугадывание будущих ненаблюдаемых величин, то в контексте системы управления доходами оценка - подборка параметров прогнозной модели, используется не часто, а вот прогнозирование – использование оцененной модели для предугадывания значений – используется часто.

Выбор подходов и методов для оценки и прогнозирования довольно обширный. Самые распространённые непараметрические методы прогнозирования – это непараметрический метод наименьших квадратов с оцениванием точности прогноза, метод Эфрона, метод Виллемейна и другие. А методы оценивания – это непараметрические методы доверительного оценивания точки наложения (встречи) двух временных рядов, непараметрических оценок плотности распределения и другие. Некоторые ученые доказали, что если улучшить хотя бы на 10 процентов точность прогноза спроса, то доход может увеличиться от 0.5 процентов до 3 процентов.

Цензурирование спроса является одним из самых сложных вопросов при прогнозировании спроса в системе управления доходами, так как имеются различная продукция, которая производится в определённом количестве и при если будет полностью выкуплен ассортимент, это не будет значить, что спрос имеет значение ровно столько, сколько было продано. Исторические данные же наоборот, должны отображать только реальный спрос, а вместо этого отражают лишь фактические продажи, что не является наилучшим показателем в прогнозировании спроса.

Использование цензурированных данных может привести как к переоценке, так и к недооценки спроса, а это в свою очередь может очень сильно повлиять на доход, то есть привести к потере дохода. Чтобы избежать этой проблемы, сначала нужно оценить параметры распределения общего спроса по цензурированным данным и только потом проводить прогнозирование. Процедура оценивания параметров распределения общего спроса по исходной цензурированной выборке имеет название восстановление общего неограниченного спроса, а методы называются – методы восстановления.

Очень сложно измерить общий неограниченный спрос, но многие исследователи утверждают, что если его использовать, можно сделать прогноз более точным и следовательно поиметь большую прибыль.

**2.3 Цензурирование спроса. Основные понятия и примеры**

Можно рассмотреть простейшую схему управления доходами на примере продавца и продуктов питания. Продукты питания имеют свойство портиться, что-то портится за день, что-то за неделю, и в этой ситуации продавцу нужно установить цену так, чтобы за период продаж продукты не испортятся, а покупатели в полной мере смогут купить все те продукты, что пожелают. Если поставить слишком низкую цену продуктам, то они будут быстро раскуплены, но доход от них будет крайне мал или вообще будет отсутствовать. При этом, так как все продукты будут раскуплены, а новые приехать не успеют, вновь пришедшие покупатели не смогут получить желаемый продукт и тем самым, возможно, больше никогда не придут в этот магазин. Если же поставить цену слишком высокую, то есть вероятность, что многие продукты вообще не будут куплены, так как какие-то покупатели решат купить аналогичный продукт за меньшую цену в другом магазине, и оставшиеся не проданные продукты будут попросту выкинуты, так как продавать испорченные продукты нельзя. Следовательно, продавец потеряет очень много прибыли. В этом и заключается суть, чтобы понять, какую именно цену нужно выставить на продукты, чтобы извлечь наибольшую выгоду для продавца.

Решение продавец будет принимать на своём опыте и своём прогнозе на спрос. Ему требуется спрогнозировать количество востребованных продуктов на разных ценовых уровнях и в последствии регулировать цены для большой эффективности. Пока время идёт, ещё не купленные продукты начинают портиться и наступает точка, когда нужно либо быстро продать, либо придётся выбрасывать. Продавец может выставить скидку на эти продукты, за частую в наших магазинах так это и происходит. Если посмотреть на продукты со скидкой и взглянуть на их срок годности, можно увидеть, что срок истечёт через короткий срок. Тем самым вроде и покупателю хорошо, он купил хороший продукт с меньшей ценой, и продавец не потерял прибыль полностью. Если же товар не продаётся так быстро, как ожидал этого продавец, он может снизить цену, чтобы извлечь наибольшую прибыль. Но намного сложнее что-либо сделать, когда товар наоборот уходит к покупателям очень быстро и, возможно, его даже не хватает. Тогда продавцу труднее сделать выбор, так как сегодня товар ушёл быстро, а завтра – неизвестно. И что было бы, если бы у него ещё больше было товара – не понятно. И в этом смысле, данные продавца являются цензурированными.

Пример продавца является простейшей иллюстрацией схемы работы управления доходами и её проблему оценивания общего спроса цензурированных данных. Таким образом, можно, да и даже нужно, воспользоваться определёнными подходами, чтобы преобразовать цензурированные данные в более точные оценки общего неограниченного спроса. Множество методов существует для восстановления данных по цензурированным данным. Есть простые, есть сложные, которые подразумевают полную оценку вероятностного распределения.

**2.3.1 Определение цензурированного спроса.**

Процесс сбора исторических данных может быть собран как с истинными данными, так и с цензурированными. С истинными данными происходит тогда, когда сколько людей приобрело тот или иной продукт, столько в историю данных и записалось, цензурированные данные получаются тогда, когда в какой-то момент времени продукция закончилась, а люди продолжали приходить и желали приобрести товар, но, так как у них не удавалось этого сделать, в исторические данные попадал лишь тот максимум, что был приобретён.

Исходя из этого, спрос является цензурированным, если спрос на товар больше, чем бронированное количество товара в рассматриваемый момент времени. Это можно проиллюстрировать в таблице формирования цензурированной выборки данных по спросу. В этой таблице есть три строки: Общий спрос – количество людей, желающих приобрести товар, предел бронирования – количество имеющегося товара, реальные продажи – те продажи, что получилось продать. Серым закрашены те ячейки, где реальные продажи являются цензурированными из-за того, что продать больше нет возможности, так как имеющегося товара меньше, чем общего спроса.

Таблица 1. Пример формирования цензурированной выборки для данных спроса

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Общий спрос | 12 | 11 | 9 | 18 | 7 | 20 |
| Предел бронирования | 9 | 20 | 10 | 17 | 5 | 25 |
| Реальные продажи | 9 | 11 | 9 | 17 | 5 | 20 |

**2.3.2 Определение предела бронирования и уровня защиты**

Главной задачей систем управления доходами является увеличение и максимизация прибыли за счёт правильного распределения имеющихся ресурсов. Это делается благодаря пределов бронирования и уровням защиты.

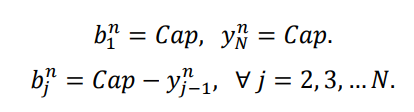
Уровень защиты является одним из самых главных параметров, который характеризует стратегию продаж по ценовым классам.

Существует два типа уровни защиты – это невложенный и вложенный. Невложенный уровень подразумевает, что каждому ценовому классу соответствует определённый сегмент в общей вместимости, который можно продать только покупателям этого класса, так как продавать высшим по цене класса становится невыгодно из-за того, что позиция будет закрыта сразу же после приобретения всех элементов выделенных под этот сегмент, несмотря на то, проданы ли места в остальных классах или нет.

Вложенная структура подразумевает, что уровень защиты Yj определяется для совокупности классов, где начиная с j и выше, и есть количество продуктов, зарезервированных для продажи определённым классам. Уровень защиты для самого нижнего класса является количеством всех продуктов, так как все продукты в сумме и будут соответствовать ей.

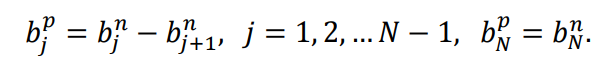
Предел бронирования так же характеризует стратегию продаж по ценовым классам. При заданной вместимости продукции в магазине (Cap), предел бронирования bj для определённого ценового класса j, является максимальная количество продукции, которые продавец готов продать за определённую цену определённому классу. После продажи этот класс закрывается. При невложенной структуре предел бронирования является таким же, как и уровень защиты. При вложенной структуре, предел бронирования и уровень защиты образуют вложенную иерархическую структуру. Предел бронирования является полной вместимостью для самого высокого класса. При такой структуре не возникает таких проблем, как когда более высокие классы уже раскуплены, а нижние ещё имеются.

Для вложенных классов предел бронирования и уровень защиты связаны соотношением:



Вложенный предел бронирования (и уровень защиты ), зная эти параметры, можно определить минимальное и максимальное количество продуктов, которые стоит зарезервировать на основе прогноза спроса в этом классе.

Предел бронирования для невложенных и вложенных ценовых классов связаны соотношением:



**2.4 Вывод**

В этой главе была рассмотрена одна из наиболее важных прикладных областей исследования операций, а именно управление доходами. Управление доходами – это особые тактики и подходы, которые позволяют предугадать поведение потребителей на определённом рынке и оптимизировать ценовую политику и доступность определённого вида товаров для потребителей, с целью максимизации прибыли.

В сфере услуг всё чаще применяются тактики и подходы управления доходами. Для этого используются современные компьютерные технологии, целые штабы аналитики и сбора данных, и прочие другие не менее важные составляющие. В этой главе разобрана общая схема работы системы управления доходами. Описаны главные шаги для успешного функционирования, в частности:

1. Сбор исторических данных
2. Оценка и прогнозирование имеющихся данных.
3. Оптимизация полученных данных
4. Контроль тех данных, что получили на предыдущих шагах.

Один из самых важных составляющих блока оценки и прогнозирования является процесс восстановления спроса по цензурированным данным.

Цензурированные данные возникают из-за того, что реальный спрос может быть намного больше, нежели реальное количество продукции. Количество продукции может быть меньше реального спроса, так как на продукцию установлены ограничения, а именно – пределы бронирования. Когда достигается этот предел, продажи прекращаются и как следствие, информация о реальном спросе теряется. Если не брать этот момент во внимание, последующие попытки спрогнозировать спрос по историческим данным окажется неверным или же будет иметь большую ошибку. Из этого вытекает, что продавец или производитель будет терять прибыль, что не есть хорошо. Эту проблему позволяет решить процесс восстановления спроса цензурированных данных. Таким образом, методы, позволяющие восстановить цензурированные данные обладают высокой значимость, так позволяют увеличить точность прогноза, а вместе с ней и максимизировать прибыль.

**3. Непараметрические методы восстановления распределений спроса**

**3.1 Математическая постановка задачи восстановления спроса**

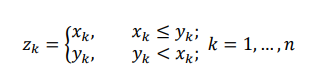
Имеется случайная величина Х, её функция распределения равна Fx(a1,…,an), a1,…,an – параметры распределения. Данная переменная описывает общий спрос. Так же есть выборка из n наблюдений случайной величины 𝑥̅ = (𝑥1,… 𝑥𝑛).

Случайная величина Y является ограничением этого спроса. Функция распределения 𝐹𝑌(𝑏1, … , 𝑏𝑠).

X и Y независимые случайные величины и они могут подчиняться как непрерывным законам распределения, так и дискретным.

Так же рассмотрим случайную величину, которая описывает величину продаж: Z = min(X, Y)

Тогда z1 = min(x1, y1), …, zn = min(xn, yn), где = (z1, …, z2) – является цензурированной выборкой, которая построена на основе выборки случайной величин X. Если xk ≤ yk, то значения являются нецензурированными, если же yk < xk, тогда yk = xk и xk будут цензурированными.



То есть, если взглянуть на задачу восстановления общего неограниченного спроса по цензурированной выборке с математической точки зрения, то смысл состоит в оценке параметров распределения a1, …,an случайной величины Х по цензурированной выборке = (z1, …, z2).

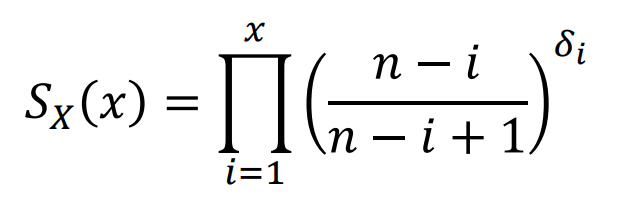
В дальнейшем может помочь информация:

1. Известен закон распределения случайной величины Х
2. Для некоторых значений известно, цензурированы данные или нет. То есть имеется какой-либо индикатор или флаг, который отвечает за цензурированность данных.

**3.2 Метод Каплана-Мейера**

Данный метод является одним из самых известных. Он обладает асимптотической эффективностью, что помогает наилучшим образом оценивать результат.

Существует упорядоченная выборка = (z1, …, z2) из n наблюдейний. Так же существует выборка = (), которая содержит индикаторы цензурирования. S(x) = 1 – F(x) – является оценкой функции надежности, которая рассчитывается по методу Каплана-Мейера следующим образом:



Здесь Sx(x) – оценка функции надежности, n – общее количество наблюдений, – это произведение по текущим наблюдениям, которые не превосходят x, а j – это индекс наблюдения.

В основном, этот метод применяется, чтобы оценить закон распределения, тем самым он является неким аналогом эмпирической функции распределения, которая применяется к цензурированным значениям. Так же этот метод можно применить только к тем случаям, где известна полная информация об индикаторах цензурирования.

Этот метод очень интересный и, возможно, ему стоило уделить большего внимания, но тема диплома непосредственно состоит в методе, что рассматривается следующим. Этот метод называется: метод бутстрап.

**3.3 Методы бутрстрап.**

Такая методика, как бутстрап, была впервые предложена американским статистиком Брэдли Эфроном в 1977 году. Вся суть метода состоит в том, чтобы построить эмпирическое распределение, то есть получить приближение теоретической функции распределения, которое построено с помощью выборки из него. Делается это с помощью многократной генерации выборок из исторических данных, которые помогают смоделировать симуляцию неизвестного распределения на основе бутстрапированных выборок.

Этот метод очень хорошо себя показывает, если данные являются нерегулярными, то есть имеют какие-то провалы. Приведём пример алгоритма данного метода на примере продаж:

Обычно продажи рассчитываются по месяцам, в году всего 12 месяцев и поэтому из ряда можно взять случайных 12 значений. На самом деле можно брать любое значение, но от количества взятых значений может сильно подняться или опуститься конечная ошибка. Возьмём эти 12 случайных значений около 10 тысяч раз. Опять же, можно брать и больше, и меньше. Но логично было бы предположить, что чем больше количество повторений, тем больше вероятность сделать более точный прогноз. Далее, по каждому из этих 10 тысяч рядов, в которых лежат случайные значения, рассчитывается среднее значение каждого такого ряда. А по получившемся средним значениям этого ряда, считаем ещё раз среднее значение. Таким образом в конечном итоге получается средние месячные продажи.

Так же можно выделить фактическое распределение, с помощью которого можно повысить точность прогноза. Делается это так:

1. Считаем среднее значение по всем 10 тысячам рядам выборки
2. Далее сгруппируем ряд по этим средним
3. Выделим важные показатели, такие как: количество, среднее, максимум, минимум, сумма и так далее.
4. По этим выведенным параметрам делаем прогноз по вышеприведённому алгоритму.

Существует блочный бутстрап и бутстрап максимальной энтропии.

**3.3.1 Блочный бутстрап.**

Суть блочного бутстрапа скользящих блоков состоит в том, чтобы из выборки выбирать не отдельные значения, а группу значений. Выборка для блочного бутстрапа формируется следующим образом:

1. X = {X1,…, Xn} – выборка наблюдаемых данных.
2. Для некоторой заданной величины L∈[1,n) формируются блоки

Bi = (Xi, …,Xi+L-1), где i ∈[1,n-L+1)

1. Из распределения всех величин i выбирает b количество блоков, чтобы bL было не меньше длины реальной выборки.

**3.3.2 Бутстрап максимальной энтропии**

Суть бутстрапа максимальной энтропии состоит в том, чтобы построить определённое количество точек в окрестностях искомых данных так, чтобы конечная выборка максимально походила на искомую. Формируется следующим образом:

1. X = {X1,…, Xn} – выборка наблюдаемых данных.
2. X1,…, Xn упорядочены по возрастанию и на каждом интервале оценивается плотность максимальной энтропии.
3. Из равномерного распределения на интервале [0,1] случайно выбираются значения случайных величин, которые не превышают максимальное значение в выборке(квантилей).
4. Полученную выборку упорядочиваем.

Так же данные методы можно видоизменить, улучшить или сделать их более пригодными к тем или иным данным. Например:

1. Метод скользящего среднего с блочным бутстрапам, где L=30.
2. Метод скользящего среднего, который перекрывает блоки в симулированных рядах с усреднением. Новые ряды формируются в соответствии с нижеприведённым алгоритмом:
   1. Интервал в L = 30 между точками блока, равен t дневных изменений. t ∈ [10;30)
   2. Для тех дней, для которых существует более одного значения изменений, конечное значение рассчитывается методом простого среднего
3. Метод скользящего среднего, который перекрывает блоки в симулированных рядах со случайным средним. Новые ряды формируются так же, как и в методе выше, только усреднение значений происходит по формуле x=z1z2x1+z2(L-z1)x2+(L-z2)x32. Здесь z1, как и z2 берутся из независимого равномерного распределения в диапазоне [0;1]. Если же x3 отсутствует, то z2 принимает значение 1. А х1,х2 и х3 соответствуют одной и той же симулируемой точке.
4. Метод скользящего среднего, который перекрывает блоки в симулированных рядах со случайным выбором. Новые ряды формируются так же, как и в методе выше, только для всех n значений, которые соответствуют одному дню, все значения z берутся из равномерного распределения в диапазоне [0;1], а значение x равняется одному из значений x1,x2,x3 в том случае, если z находится в диапазоне [0;1] с шагом 1/n.
5. Бутстрап максимальной энтропии используемый для фактических значений. Этот метод оставляет связь между наблюдениями и при этом никак не ограничивает значения симулированной выборки.

**3.4 Сглаживание цензурированных данных.**

В продажах иногда присутствует такая продукция, которая продаётся очень редко или продаётся в разное время и по непонятным причинам. Например, какая-нибудь специализированная запчасть для автомобиля, причём для какого-нибудь редкого автомобиля. У такой запчасти могут быть продажи очень «порванными». То есть за месяц её могут купить всего пару раз, а может, даже и ни разу. Если брать годовую статистику по таким продукциям, можно увидеть, что её продажи оборваны. Это продемонстрировано на Рисунке 2.

Рис. 2. Пример оборванных данных.

На рисунке 2 продемонстрированы оборванные продажи, где показано 12 месяцев и продажи в каждом из этих месяцев. Как видно по этой гистограмме, есть месяцы, где продаж вообще не было, или же есть месяцы, где уровень продаж выше, чем в совокупности в нескольких других месяцев. Если в таком случае использовать традиционные методы прогнозирования, то есть большая вероятность ошибки, так как при таких данные традиционные методы являются неэффективными. А если использовать простейшие описательные статистики, такие как: минимум, среднее, максимум, это приводит к ненадёжным оценкам, что приводит к тому, что покупатели либо отказываются, либо наоборот вообще отсутствуют, то есть деталь была заказана в слишком большом количестве. Но из-за того, что делать с такими данными что-то нужно, учеными были придуманы некоторые способы прогнозирования таких продаж. Один из таких способов – это сглаживание продаж.

Сглаживание продаж – это искусственное дополнение или занижение данных, в частности в тех местах, где данные отсутствуют. А вот о методах сглаживание дальше пойдёт речь.

**3.4.1 Простое сглаживание**

Самый примитивный и простой в реализации способ – это способ простого сглаживания. Его алгоритм приведён ниже:

1. Сортируем все данные
2. С вероятностью Р вытаскиваем пример из данных
3. С вероятностью 1 – Р берём случайный интервал между точками среди данных и от туда берём равномерно распределённое число
4. Вставляем это число

Где Р – это вероятность появления старых точек по данным.

Анализ и сравнение данных с помощью данного метода будет рассмотрено в следующей главе. Сейчас же перейдём на более «умные методы».

**3.4.2. Метод Кростона**

Самая главная проблема при прогнозированные в таком случае, когда отсутствуют показатели продаж, это то, что не учитывается характерная особенность рядов, которые рассматриваются, при этом нулевые значения рассматриваются как обыкновенные наблюдения.

Решение данной проблемы в 1972 году предложил Кростон. В принципы прогнозирования заложено, что рассматриваются два объекта, это вероятность, что спрос будет ненулевым и величину ненулевого спроса. С помощью схемы Бернулли моделируется вероятность независимого события появления нулевого спроса. Благодаря этому, появляется предположение, что распределение спроса имеет примерно вид, как на рисунке 3

Рис. 3. Модель Кростона: распределение спроса

Такой подход оценивает три параметра:

1. Вероятность, что спрос будет отсутствовать.
2. Средний спрос за период
3. Дисперсию спроса

С помощью простого экспоненциального сглаживания будут оценены первые два параметра: вероятность отсутствия спроса и средний спрос за период, а дисперсия спроса будет оценена по тем значениям, что являются ненулевыми. Исследования огромных временных рядов привело к нескольким выводам по поводу эффективности данного метода, например, что такой метод является более эффективным, нежели простое экспоненциальное сглаживание, при прогнозировании среднего спроса за период. Так же было выявлено, что ни метод Кростона, ни метод экспоненциального сглаживания, не являются эффективными для прогноза совокупного спроса за период.

Приняв во внимание, американский математик Виллемейн в 1994 году предложил свой подход, основываясь на методе Кростона. О нём мы поговорим дальше.

**3.4.3 Метод Виллемейна**

Суть метода Виллеменйа состоит в том, что можно моделировать нормальное распределение на самими ненулевыми значениями, а их логарифмами.   
Это позволяет избежать ситуации, когда появляются неотрицательные значения изучаемых рядов. Алгоритм в простейшем виде выглядит так:

1. Берём одно случайное значение x из исторических данных
2. Заменяем это число на x + \* + 1

Если значение данного уравнения будет меньше нуля, то оставляем искомое значение, иначе заменяем.

Как говорилось выше, данный метод хорош в тех случаях, когда данные порваны и, если мы наблюдаем ряд продаж, где присутствуют нулевые решения, воспользоваться им можно путём процедуры бутстрапирования. То есть, например, у нас есть данные за последний 24 месяца и нам нужно оценить спрос на ближайшие три месяца. Из всего ряда случайным образом берутся три произвольных значения, считается сумма этих значений, запоминается и процедура повторяется снова n-ое количество раз. После окончания подсчётов появляется возможность наиболее точно дать оценку рациональному запасу продукции и получаем оценку распределения совокупного спроса. Если же к этому добавить сглаживание по вышеописанному алгоритму, то может получиться довольно неплохой анализ, который может позволить наиболее точно спрогнозировать число продаж и тем самым продавец сможет запастись наиболее верным количеством продукции, что в конечном итоге приведёт к максимизированию прибыли.

Но, очень важно понимать, что данные, для работы с этими методами, должны быть не цензурированными. То есть там, где нули, значит, действительно не было продаж, а не попросту не было товара на складе. Ведь главный смысл понять и предугадать наиболее точный спрос.

**3.5 Вывод**

В данной главе были рассмотрены основные методы восстановления спроса по цензурированной выборке, такие как: метод бустрап, метод Виллемейна, методы сглаживания и другие. Так же, в главе представлена математическая постановка задачи восстановления спроса по общей цензурированной выборке.

Это глава позволяет нам перейти к следующему пункту дипломной работы, а именно реализация методов, работа с данными, обработка данных, получение оценки и другое.

**4. Практическая работа**

Во всех главах выше было теоретическое описание вопроса**.** Было выяснено насколько важна область прогнозирования цензурированного спроса и рассмотрены методы, которые позволяют получить наиболее высокие оценки для последующей максимизации прибыли.

Выше уже было сказано, что для действительно качественной реализации нужно иметь качественные данные и, в добавок, нужно иметь какие-то сторонние приложения или же сторонние программное обеспечение для прогнозирования спроса. Но если брать какие сторонние приложения, это ставит нас в рамки и тем самым мы не сможем исследовать тот или иной метод, а значит, не сможем его как-то улучшить, преобразовать или вообще придумать новый на основе старого. Поэтому, никаких сторонних программных обеспечений использовать для прогнозирования спроса в данной дипломной работе не будет. Вместо чужого программного обеспечения для достижения наших целей, мы будем пользоваться языком программирования python. Его стандартными механизмами и библиотеками.

Так же, очень важный вопрос – это данные. Данные нужно где-то взять, проанализировать их, разобрать, и на их примере показать работоспособность того или иного метода. Данные будут естественно взяты с внешнего источника, так как найти свои данные будет довольно-таки проблематично.

Далее мы рассмотрим сам язык программирования python, узнаем, как с помощью него обрабатывать данные, рисовать всякого рода таблицы, гистограммы и прочее. Рассмотрим сами данные, откуда они были взяты, что они обозначают и на примере языка python реализуем методы прогнозирования спроса. С помощью реализованных методов прогоним наши данные, выясним сильные и слабые стороны каждого метода, попытаемся улучшить данные методы какими-то своими наработками или же комбинацией одного метода с другим.

* 1. **Язык программирования python**

Почему выбор выпал именно на этот язык ясен и понятен. На данный момент существует огромное количество языков программирования и практически каждый день появляются новые. Но есть языки, которые уже устаялись в той или иной области. Например, язык программирования java очень хорошо для написания огромных кроссплатфомремнных программных обеспечений или же веб приложения, которые будут служить людям многие годы. Языки программирования C и С++ очень хороши в написании систем, которые работают непосредственно с железом. И писать на таких языках какие-то простые вещи, которые требуют математических вычислений не особо разумно. Например, в этих языках довольно сложно работать с некоторыми библиотеками и, например, то же число, у которого более 20 знаков, будет большой проблемой для реализации методов прогнозирования спроса.

По всем этим причинам был выбран язык программирования python. Данный язык имеет понятно интуитивный синтаксис, в котором легко и быстро разобраться. Так же этот язык имеет ряд библиотек, которые позволяют легко и быстро работать с огромными данными, позволяют изображать гистограммы и графики, которые отображают данные для наглядного понимания. Так же данный язык даёт возможность даже самые типичные задачи программирования, которые на тех же С и java реализуются в пять строчек, реализовать их в одно строчку. Например, чтобы пройти по массиву, а мы будем очень много работать с массивами данных, и добавить каждому элементу число один, реализация на С и java будут идентичны:

for (int i = 0; i < len(data) ;i++) {  
 data[i] = data[i] + 1;  
}

Здесь data – это массив данных, i – счётчик итератора, который идёт по каждому элементу массива, len – длина массива. Код представленный выше лишь показывает как прибавить к элементу массива единицу. А что если бы пришлось делать намного больше операций с элементам массива, например, выбирать только чётные данные, потом делать какие-то преобразования с ними, а затем ещё что-то. Согласитесь, код бы стал в разы больше. Но язык программирования python на этот счёт предлагает лаконичное решение, которое подкупило меня в выборе языка:

data = [x + 1 for x in data if x % 2 != 0]

В этом примере демонстрируется лёгкая и приятная работа с массивом, с помощью которого сразу же происходит и увеличение элемента массива на 1, и выборка по определённым данным, в данном случае берётся каждое нечётное число.

Так как у нас имеются данные, нам нужно с ними как-то работать, анализировать, отображать и прочее, в языке python есть много библиотек для работы с такими вопросами.

Например, чтобы легче воспринимать данные, люди обычно рисуют графическое представление этих самых данных, и библиотека matplotlib прекрасно себя в этом проявляет, с помощью неё можно нарисовать всё то, что нам может понадобиться для более точного и правильно анализа методов прогнозирования спроса. Например, с помощью кода представленного ниже, можно отрисовать точку с координатами (1,1) на системе координат:

import matplotlib.pyplot as plt

fig = plt.figure()

plt.scatter(1.0, 1.0)

plt.show()

Получится следующий рисунок:

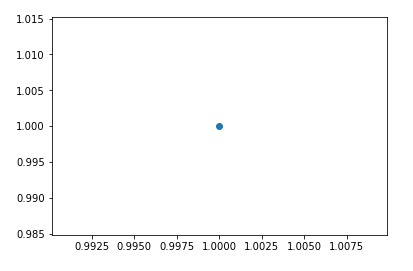


Рисунок 4. Точка с координатами (1,1) на системе координат нарисованная с помощью matplotlib.

Так же с помощью данной библиотеки можно будет рисовать более сложные вещи, поэтому мы её и выбрали.

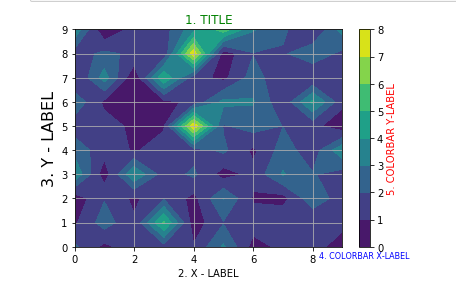
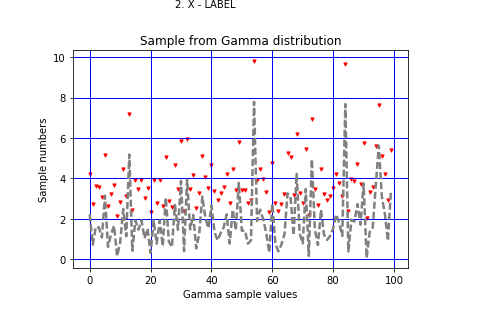
 

Рис. 5. Более сложные диаграммы нарисованные с помощью библиотеки matplotlib

А для работы с самими данными мы будем использовать такие библиотеки как pandas и numpy. Numpy является классической библиотекой, для работы с массивами данных. Она позволяет без каких-либо проблем обрабатывать данные, находить среднее, минимум, максимум и много других полезных функций, которые могут понадобится при работе с данными.

Более сложные вещи с данными позволяет делать библиотека pandas. Во-первых, эта библиотека позволяет считывать данные из разных источников, будь то excel файл или же csv файл, а может даже данные из интернета. Так как эта библиотека даёт широкий выбор как именно считывать данные, какие именно считывать данные, сразу же с ними что-то делать или не делать, мы её и выбрали.

Помимо считывания из разных источников, данная библиотека умеет работать с самими данными таким образом, что мы буквально говорим «я хочу, чтобы отрицательных значений не было в данных» и библиотека всё делает за нас, не нужно писать каких-либо отдельных методов или функций для преобразования данных. Библиотека всё сделает за нас.

По сути, работа в этой библиотеке состоит из таблиц. Логично, что данные находятся обычно в таблицах, например, есть данные по продажам за период. Логично предположить, что удобнее всего работать с этими данными тогда, когда они представлены в таблицах. То есть имеется два столбца, первый столбец - дата, второй столбец - продажи. Помимо продаж и даты могут быть ещё какие-нибудь второстепенные или же важные данные, такие как: флаг, был ли предмет возвращён, когда продажи отрицательные, был ли товар украден или ещё что-то. Помимо таких банальных таблиц, могут быть более крупные. И так как всё это хранится как таблица, то и работа с этими данными в библиотеке pandas происходит точно так же, как и работа с таблицами в SQL, оно и логично, хороший подход, который ясен многим.

Помимо всех этих полезных и удобных вещей, есть много других удобных функций. Например, код представленный ниже выдаёт среднее значение по всей продукции за февраль 2012 года:

df.loc['2012-Feb'].mean()

Это очень удобно и органично, никаких дополнительных самописных функций, никаких дополнительных излишеств, лишь библиотека и работа с данными.

Благодаря вышеуказанным плюсах, и многим другим, что не были указаны, и пал выбор на эту библиотеку и на сам язык программирования python. Всё это позволяет реализовать, наиболее точно оценить и сделать вывод по непараметрическим методам прогнозирования спроса, в частности метода бутстрапа. Но это всё лишь маленькая часть большой работы, следующая маленькая часть – это данные, о которых будет идти речь в следующей главе.

**4.2 Данные.**

Очень важную часть, в наиболее корректном прогнозировании цензурированного спроса, составляют данные. Как уже говорилось ранее, данные методы для восстановления цензурированного спроса хорошо работают с данными, которые на каких-то отрезках имеют нулевые значения. Нулевые значения означают, что в каком-то месяце, дне, недели или году, отсутствовали продажи по той или иной причине. Бывает множество различного товара, от повседневных товаров, которые покупают изо дня в день миллионы потребителей, до таких товаров, что покупают раз год и то не факт. Не обязательно, чтобы данные совершенно не покупаемыми, главное, чтобы отсутствовало такое понятие, как сезонность или что-то на него похожее. То есть, если будут скакать данные, это нам будет подходить. Вот именно на таких данных, которые имеют «просадки» и стоит тестировать написанные нами методы, о которых будет идти речь после описания данных.

Данные были взяты с сайта Kaggle.com. Этот сайт предоставляет данные для различных олимпиад, чтобы люди с помощью них учились и развивались в таких направлениях как: машинное обучение, программирование нейронных сетей, прогнозирование спроса и прочее. На этом сайте есть множество данных, но мы будем использовать одну таблицу, которая имеет название «transactions». Наименование этой таблицы означает, что в ней содержатся данные об каких-то транзакция, ну или сделках. Что именно за товар продаётся в этих магазинах неизвестно, но это не имеет особой важности, если данные нам подходят. Конечно, если бы мы знали, что именно эти магазины продают, можно было бы чисто логически придумать дополнительные условия, которые можно добавить в реализацию методов прогнозирования спроса, но так как мы не знаем, что именно продают данные магазины, довольствуемся тем, что имеем. То есть, в таблице данные о проданных товарах. Данные реальные, они были собраны с 54 магазинов Favorita, которые расположены в Эквадоре. В этой таблице содержится лишь дата продажи и количество транзакций. Идеальный вариант для работы. Сами данные выглядят примерно вот так:

Таблица 2. Пример исторических данных.

|  |  |  |
| --- | --- | --- |
| date | store\_nbr | transactions |
| 2017-08-16 | 1 | 770 |
| 2017-08-16 | 2 | 400 |
| 2017-08-16 | 3 | 872 |

где date – дата продаж, store\_nbr – номер, transactions – количество продаж.

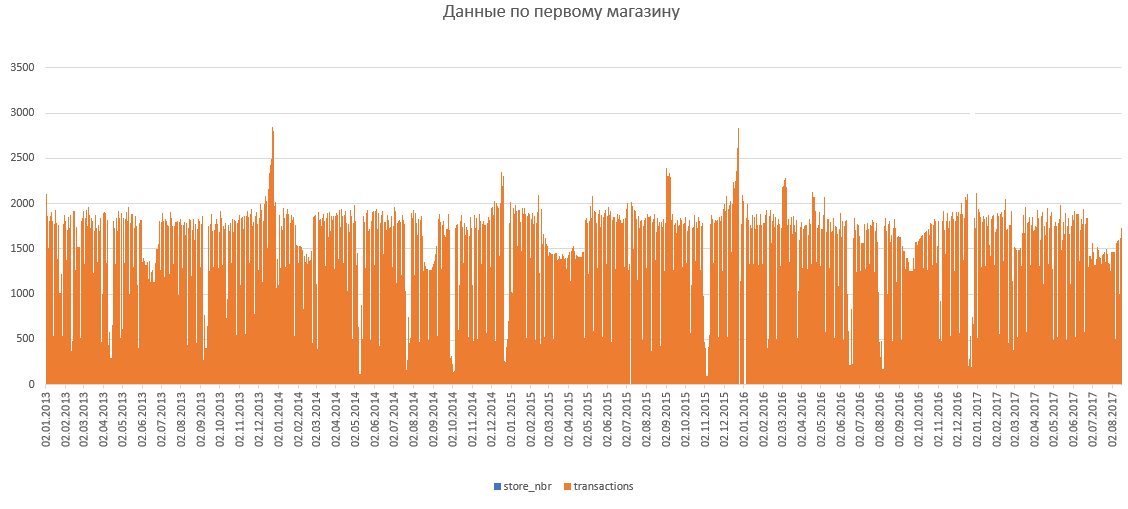
Первой датой в предоставленных данных является 1 января 2013 года – вторник. А последняя дата 15 августа 2017 года – тоже вторник. То есть мы имеем данные по 54 магазинам за период практически пяти лет. Это довольно неплохой срок, чтоб научиться на основе него делать прогнозы. Прогнозировать можно как по одному магазину, как и все вместе. Стоит показать наглядно как выглядит гистограмма данных за этот длительный период: 

Рис. 6. Пример данных по первому магазину

По данным можно заметить, что они не являются идеальными. Всё по причине того, что это данные всех совокупных продаж, а не по отдельным товарам. Но всё равно в некоторых местах можно заметить, что продажи ведут себя странно и обосновать их сезонностью нельзя. То есть они где-то проседают. Можно даже выбрать какое-то значение продаж и принять его за ноль. Например, средняя продажа 1500. Всё, что ниже, считается нулём. Тогда продажи точно будут подходить под наши методы, но делать этого мы не будем, мы будем работать с теми данными, что предоставлены.

Как уже говорилось, таких магазинов 54 и у каждого магазина примерно похожий график продаж. На каждом магазине можно заметить, что в начале 2016 года отсутствуют продажи вообще, это последствия землетрясения, которое случилось в начале 2016 года. Естественно, в эти дни магазины либо не работали, либо же раздавали свою продукцию за просто так, если она могла как-то помочь жителям страны.

Как именно мы будем работать с данными? Так как есть только исторические данные, но нет данных, по которым мы сможем сделать вывод, работает ли наш прогноз, мы немного урежем данные. То есть за основу, то есть за исторические данные, мы возьмём все данные, кроме последней недели. Последняя неделя как раз и будет являться той, по которой мы будет сравнивать. Это может быть не совсем корректно, так как, возможно, именно в эту неделю были какие-то проблемы или спад, или наоборот увеличение спроса, что в конечном итоге повлияет на оценку, но процент ошибки не должен очень сильно увеличиться. Поэтому даже с такими данными можно будет сделать выводы. К тому же, когда были описаны даты начала и даты конца данных, было уточнено, что начало – вторник и конец – тоже вторник. Поэтому, прежде чем составлять исторические данные и данные на проверку, нужно данные обрезать, чтобы с ними было проще работать по недельно или, если нужно будет, по каким-то конкретным дням недели. Поэтому мы обрезаем первые шесть дней и последние два дня. И из полученных данных получаем нужные нам данные, с которыми и будет производить последующие действия, анализ, обработка, отрисовка и другое.

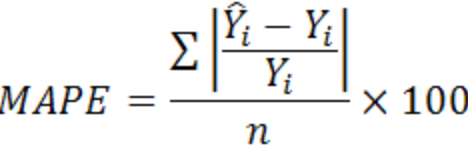
**4.3 Реализация.**

Первое, с чего всё начинается – это со считывания данных. Как уже было сказано, у нас есть таблица в формате csv, которая имеет наименование transactions.csv, в которой есть 3 колонки – дата, номер магазина и число продаж. После того, как данные будут прочитаны, мы сортируем их по дням. Далее добавляем столбец, который будет означать день недели. Это нам пригодится в том случае, если мы захотим сделать выборку по дням недели.

Начало данных – это вторник, и конец данных – тоже вторник. По этой причине, второе, что мы сделаем, это обрежем данные, чтобы начало было с понедельника, а конец был воскресеньем. Это нужно для того, чтобы правильно воспринимать и обрабатывать данные. Чтобы выборка, например, по дням, всегда строилась исключительно по тем дням, которые нам нужны.

По причине того, что у нас есть исторические данные, но нет данных, с которыми можно сравнивать, третье, что нужно сделать, это выделить из исторических данных одну последнюю неделю, которая будет являться результирующей, то есть той неделей, с которой мы будем сравнивать. В рамках языка это делается очень просто, создаётся класс TransactionHistory, в котором будет всего два поля, оба поля массивы, содержащие данные. Первое поле – learning\_history (обучающая история), второе – last\_weak (последняя неделя). Что эти поля делают и для чего они предназначены понятно из их имён. То есть после прочтения данных, их нужно как-то отсортировать по магазинам. Поэтому мы создаём словарь. Словарь в языке python – это структура данных, которая содержит ключ и значение по этому ключу. В нашем случае ключом будет номер магазина, а значение ключа будет как раз объект класса TransactionHistory, куда мы поместим заранее разделив исторические данные и последнюю неделю для сравнения. Таким образом, при прогнозировании с помощью методов, что в дальнейшем реализуем, мы будем работать исключительно с верными данными, которые относятся только к своему магазину и не возникнет ситуации, где будет работа с данными из разных магазинов. Так же благодаря этому разделению, можно будет анализировать не все магазины сразу, а выбрать один из них, который представляет наибольший интерес. Например тот, у которого слишком большая ошибка, или тот, у которого слишком маленькая. Благодаря такой возможности, можно будет в наибольшей мере оценить работу реализованным методов прогнозирования и сделать наиболее точные выводы.

После того, как были подготовлены данные для работы с ними, пора приступать к непосредственной реализации самих методов прогнозирования. Прежде чем реализовывать самые главные методы, реализуем самый простой метод, который поможет нам сравнивать работу остальных. Назовём его simple, что означает простой. Смысл этого методы состоит в том, чтобы просто отдать сумму значений предпоследней недели. Это делается потому, что мы будем прогнозировать будущий недельный спрос и сумма последней недели является ответом прогноза, а сравнивать с чем-то нужно. И самым простым способом будет либо посчитать среднее по каждой недели, либо просто взять значений предпоследней недели. Помимо суммы значений за неделю, если будет выборка делаться по определённым дням, например, вторникам, этот простой метод, может дать ответ по конкретному дню недели от понедельника до воскресенья. После того, как мы получим ответ по этому простому методу, нужно вычислить ошибку. Это мы будем делать с помощью MAPE. MAPE – это средняя абсолютная ошибка в процентах. Процентное соотношение нам очень подходит, по причине того, если выводить результат в виде графика, сразу будет понятно на сколько процентов отклоняется ответ от реальных данных. Формула MAPE представляет из себя:



Где Yi – это фактический объём продаж за анализируемый период, – значение прогнозной модели за анализируемый период, n – количество периодов.

На языке python я это реализовал вот так:

def MAPE(correct\_value, forecast\_value):  
 return abs(correct\_value - forecast\_value) / correct\_value \* 100

В языке python слово def обозначает функцию. MAPE – наименование функции, correct\_value и forecast\_value – входные параметры в функцию. Где correct\_value – сумма значений последней недели, которую мы предварительно посчитали, forecast\_value – значений метода прогнозирования. В моей реализации на языке Python нет суммы и деления на количество периодов, так как мы имеем лишь два значения, корректные данные и спрогнозированные данные.

На примере вычисления ошибки и простого методы было показано, как именно будет оценивать ошибка и с чем она будет сравниваться. Благодаря простому методу, можно будет наглядно примерно оценить работоспособность настоящих методов. А к реализации настоящих методов приступим далее.

Главный бутстраповский метод – это метод Эфрона. Его мы и будем реализовывать. В простом варианте без каких-либо модификаций данный метод на языке python выглядит так:

ITERATION\_COUNT = 100000

DAYS\_OF\_WEEK = 7

# Метод Эфрона  
def efron(data):  
 result = []  
 for i in range(0, ITERATION\_COUNT):  
 var = sum(np.random.choice(data, size=DAYS\_OF\_WEEK))  
 result.append(var)  
 return np.average(result)

В метод мы посылаем данные – data. Далее, как было рассказано в описании методы, какое-то количество раз подряд, а именно ITERATION\_COUNT (переменная отвечающая за количество повторений), мы должны рандомно взять то количество дней, которое хотим спрогнозировать, а именно DAYS\_OF\_WEEK, то есть семь дней или одну последующую неделю. После того, как мы возьмём эти семь дней, берём сумму этих семи рандомных дней и кладём в результирующий массив. Далее собираем такие данные в результирующий массив ровно столько раз, сколько делаем повторений, в нашем случае ITERATION\_COUNT. После всех этих действий, в результирующем массиве у нас лежит количество в размере ITERATION\_COUNT сумм по семь дней и среди всех этих данных мы находим среднее, что и будет являться результатом. Как только получаем результат, получаем по нему ошибку MAPE.

Поле того, как мы получили значения и по простому прогнозированию, и по прогнозированию с помощью метода Эфрона. После того, как получили ошибку MAPE, можно построить гистограмму по этим данным и посмотреть на результат, но это уже действие следующей главы, где мы будем рассматривать и анализировать результат.

Помимо этой просто реализации метода Эфрона, можно к ней добавить выборку по дням. То есть вместо того, чтобы брать рандомные дни по всем данным, можно будет брать только понедельники, или только вторники. Тогда можно будет спрогнозировать спрос на конкретный день недели. Чтобы этого достичь, прежде чем отправить данные в метод, нужно будет из них сделать выборку по дням. Как обговаривалось ранее, библиотека pandas в составе языка программирования python умеет работать с данными как с таблицами работает SQL, поэтому одной командой из тех данных, что мы когда-то подготовили, выбираем только по определённому дню и отправляем в метод Эфрона. При этом сам метод должен как-то понять, что это выборка по дням, тк если не сделать никаких модификаций, он будет искать на семь последующих дней значения и то в сумме. Что приведёт к большой ошибке. Единственное, что изменится в методе, это появится флаг is\_day, который отвечает за то, за всю неделю или за один день данные, и при занесении в результирующий массив полученные рандомные дни мы будем делить на количество дней в неделю, то есть на семь. Тем самым мы получим конкретное прогнозируемое значение за конкретный день недели.

Помимо простого метода прогнозирования и метода бутстрапа, реализованы так же методы сглаживания. Сначала речь пойдёт о простом варианте сглаживания.

Простой вариант сглаживания представляет из себя следующий алгоритм:

1. Сортируем данные.
2. Высчитываем вероятность. Так как очень большой разброс чисел, делать будем это по диапазонам. Примерный диапазон чисел будет рассчитывать по формуле (max-min)/len(data). Далее считает вероятность выпадения именно по этому диапазону, а не по отдельным числам.
3. Вытаскиваем по вероятности числа из data.
4. Берём интервал с вероятностью 1 - p и заполняем a + (b-a) \* r, где r – рандомное значение в диапазоне от нуля до единицы.

Реализация на языке питон выглядит так:

def smooth(data):  
 copy\_array = np.copy(data)  
 copy\_array.sort()  
 data\_range = get\_data\_range(copy\_array)  
 probability = get\_probability(copy\_array, data\_range)  
 for key in probability:  
 if list(probability)[len(probability) - 1] == key: break  
 a, b = get\_interval(copy\_array, key)  
 data.append(round(a + (b - a) \* random.random()))  
 return data

Для начала копируем данные в отдельный массив, чтобы работать с копией и тем самым не испортить настоящие данные. Затем сортируем и получаем диапазон для расчёта вероятности с помощью метода get\_data\_range. После того, как мы посчитали диапазоны, по этим диапазонам нужно найти их вероятность выпадения. То есть ту вероятность, с которой может выпасть то или иное число из конкретного диапазона. Высчитываем вероятности на каждый диапазон с помощью метода get\_probability. Он нам вернёт словарь с ключом – диапазон, значение – вероятность. А дальше мы идём по этим диапазонам, получаем какой-то интервал среди отсортированных данных по этому диапазону. То есть находим в данных такой промежуток, в котором находятся значения диапазона и вставляем туда значение равномерного распределения по этому диапазону. В конечном итоге, если совместить метод бутстрапа и добавить к нему это простое сглаживание, данные должны стать более сглаженными и тем самым результат должен быть оказаться более правдоподобным, а тем самым он сможет максимизировать прибыль.

Далее, разберёмся с методом Виллемейна. Реализация на языке python данного метода выглядит так:

#### def willemain(data): for i in range(len(data) // 3): index = random.randint(0, len(data) - 1) x = data[index] data[index] = x + norm(x) \* math.sqrt(x) + 1 return data

Сразу возникает вопрос, зачем длину данных мы делим на три. Суть метода Виллемейна состоит в том, чтобы сгладить не все числа, а только какую-то их часть и мне показалось, наиболее рациональным выходом из данной ситуации, это подправить только треть данных, а не все или даже половину. Так как половина – это уже много, а четверть или ещё меньше, слишком незначительно повлияет на конечный итог. Поэтому треть – самый оптимальный выход. Тем самым, мы идём по данным и выбираем из них случайным образом какое-то число, после того, как мы его выбрали, мы его заменяем на формулу x + \* + 1. Где x – само значение из данных, – это нормальное распределение с математическим ожиданием 0 и дисперсией 1. Далее возвращаем эти данные и по ним, и с помощью методов бутстрапирования, получаем конечное значение, по которому будем искать ошибку и делать прогноз.

**4.4 Вывод**

В данной главе были рассмотрены основные принципы работы и их реализация. Сначала мы быстрым взором взглянули на язык программирования python, на его возможности и преимущества. Обосновали выбор этого языка программирования для реализации данной дипломной работы. Так же рассмотрели библиотеки, которые используем для отрисовки графиков и, которые используем для работы с огромными массивами данных, для более продуктивного прогнозирования.

Далее были рассмотрены сами данные, так как данные являются очень важной частью в прогнозировании спроса. Было выяснено, что имеется 54 магазина в Эквадоре, показано распределение данных на одном примере одного из магазинов.

Затем были рассмотрены и показаны реализации методов прогнозирования. Были показаны простой метод прогнозирования, метод бутстрапа Эфрона. И методы сглаживания: простое сглаживание, сглаживание Виллемейна.

Какие-то отдельные методы бутстрапа не были реализованы и рассмотрены, так как они являются лишь модификациями основных методов. Так же не был рассмотрен метод Кростона, по причине того, что основной темой диплома является метод бутстрап.

Так как с основной работой знакомство состоялось, осталось лишь проанализировать и обосновать полученные результаты. Этому и будет посвящена следующая глава.

**5. Сравнение и анализ оценок, полученных с помощью методов прогнозирования и сглаживания данных**

**5.1 Методы прогнозирования без сглаживания**

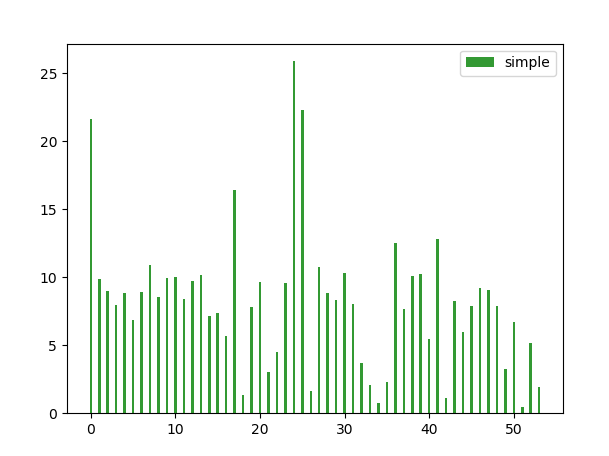
Первый метод, который будет рассмотрен и оценен, по которому можно будет сделать первые выводы, будет просто метод прогнозирования. График его работы показан ниже:

Рис. 7. График ошибки простого метода прогнозирования

На предоставленном 7 рисунке можно увидеть, что примерно во всех магазинах один и тот же уровень ошибки, он находится где-то на уровне 8.2 процентов. Но есть магазины, которые выделяются либо очень большой, либо очень маленькой ошибкой. Эти магазины стоит исследовать, чтобы понять с чем связаны такие большие или наоборот, маленькие, ошибки.

Магазины, у которых ошибка больше 15 процентов – 1, 18, 25 и 26.

Магазины, у которых менее 2 процентов – 19, 27, 35, 43, 52.

Таблица 3. Сравнительные данные по магазинам с большим и маленьким процентом ошибки.

|  |  |  |  |
| --- | --- | --- | --- |
| № магазина | Спрогнозированное значение | Значение сравниваемых данных | Среднее значение по всем данным |
| 1 | 10890 | 8951 | 10672 |
| 18 | 9098 | 7817 | 9263 |
| 25 | 6692 | 9028 | 6578 |
| 26 | 4171 | 3411 | 4441 |
|  |  |  |  |
| 19 | 8226 | 8117 | 8806 |
| 27 | 11695 | 11509 | 10375 |
| 35 | 4957 | 4920 | 4693 |
| 43 | 10585 | 10707 | 9128 |
| 52 | 16887 | 16808 | 16328 |

По таблице можно заметить, что у тех магазинов, у которых очень большой процент ошибки, прогнозируемое значение приближено к среднему значению по всем данным, а там, где очень маленький процент ошибки, прогнозируемое значение, наоборот, довольно сильно отклоняется от среднего значения по всем данным, но при этом очень близко к реальным данным. Всего такие магазинов, которые ведут себя не так как все – 9, к ним можно прибавить ещё пару с относительно низким процентом ошибки и относительно высоким процентом ошибки. Получится около 17 магазинов, что является третью всех магазинов. Чтобы как-то проанализировать то, что получилось, нужно посмотреть на результаты остальных методов. Если при остальных методах останется такая же тенденция, что примерно те же магазины имеют маленький процент ошибки, и те же магазины имеют большой процент ошибки, тогда дело в последней недели, она как-то выбивается из общих данных. Если же, будет существенная разница, то это может говорить о том, что тот или иной метод не учитывает какие-то свойства данных и не корректно с самими данными работает.

Вторым методом будет – метод бутсрапа, метод Эфрона. Результаты этого метода графически будут изображены вместе с простым методом для того, чтобы было легче сравнивать. Результаты методом Эфрона:

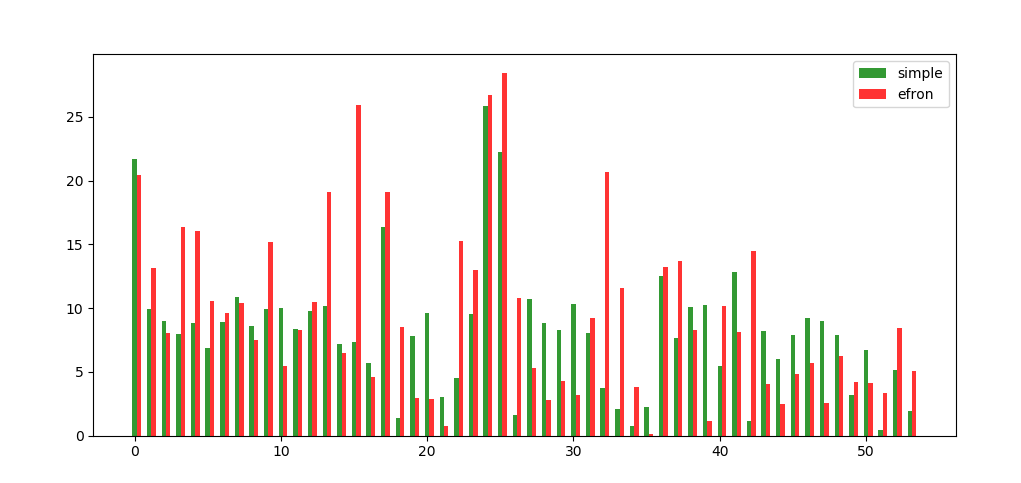


Рис. 8. Метод Эфрона и простой метод

Если проанализировать метод Эфрона, можно увидеть интересную вещь. В 26 магазинов, то есть почти в половине случаев процент ошибки уменьшился, практически во всех случаях процент уменьшился не очень сильно, где около 2-3 процента. Что является уже неплохим результатом. Но, так же, в 24 случаях – ошибка увеличилась, где-то, по сравнению с простым методом, очень сильно, около 10 процентов.

Если рассматривать этот метод без простого, то получатся следующие цифры: 9.7 процентов – средняя ошибка по всем магазинам. То есть почти на полтора раза больше, чем у простого метода. Это происходит за счёт того, что появилось целых 11 магазинов, где ошибка больше 15 процентов.

Так же мы видим, что такие магазины, как 1, 18, 25 и 26 по-прежнему имеют очень большую ошибку, это говорит о том, что последняя неделя является нестандартной, по сравнению со всеми остальными данными по этим магазинам. Поэтому, в дальнейшем мы не будем рассматривать эти магазины, хотя в данных они у нас будут присутствовать и будут немного «портить» среднее значение по всем ошибкам.

В целом, метод прогнозирования бутстрап показал себя немного хуже, чем простой метод. Но у этого метода есть наработки, которые призваны улучшить конечный результат. Эти наработки называются - сглаживанием и они будут рассмотрены далее.

**5.2 Методы прогнозирования с сглаживанием.**

Первое сглаживание будет - простое сглаживание. После простого сглаживания будет метод Виллемейна. Суть простого сглаживания состоит в том, чтобы в данные вставить примерно похожие значения, дабы в конечном итоге получить более ровный конечный ответ. Данный метод использовать вместе с простым методом прогнозирования не имеет смысла, так как простой метод содержит лишь одно число – это сумма значений предпоследней недели. Здесь просто нечего добавлять и изменять. Поэтому данный метод будет тестироваться непосредственно вместе с бутсраповским методом, так как он с лёгкостью позволяет изменять и добавлять данные. Как и в прошлом примере, данные будут показаны вместе с простым методом прогнозирования:

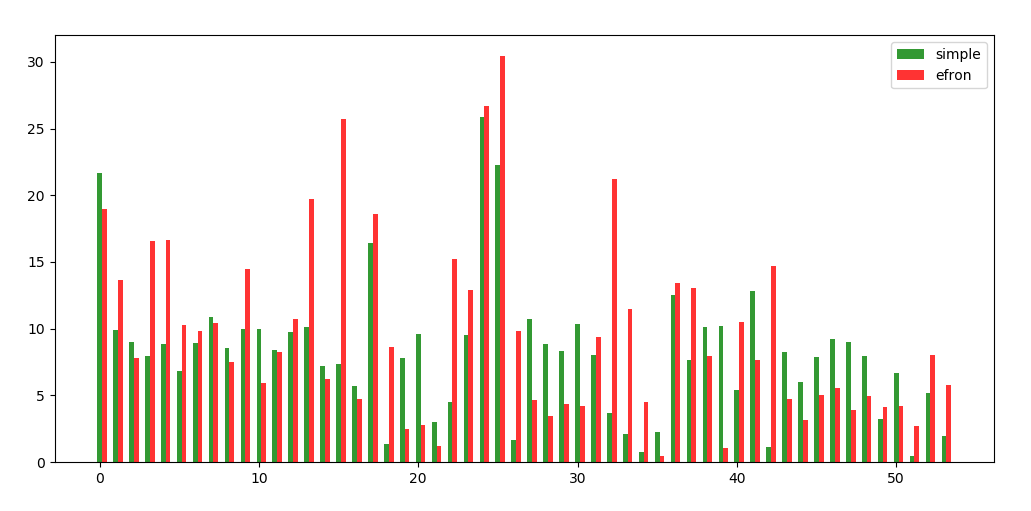


Рис. 9. Метод простого сглаживания

Метод простого сглаживания улучшил результат метода бутстрапа, но не на много. Средняя значение ошибки по магазинам составило 9.6 процентов. То есть всего на 0.1 процент меньше. Как и в прошлые разы, на тех же местах, где была большая ошибка, как и на тех же, где была маленькая ошибка, ошибки там так и остались. То есть примерно, ничего особо не изменилось, что с простым сглаживанием, что без него. Но что будет, если взять метод Виллемейна и попробовать сгладить вместе с ним.

Суть метода Виллеймена состоит в том, чтобы заменить полученные данные на более универсальные. После применения метода Виллеменйна получится следующий результат, который представлен на рисунке 10. По результатам можно увидеть, что данный метод, тоже не особо сильно изменил ситуацию, но средний процент ошибки по всем магазинам уменьшился до 9.4 процентов, что уже довольно неплохо, но всё равно чуть больше, чем в раз больше, чем простой метод прогнозирования. Это может говорить о том, что в данных, на последнее периоды идёт спад продаж, этим можно легко объяснить то, что предпоследняя неделя примерно похожа на последнюю.

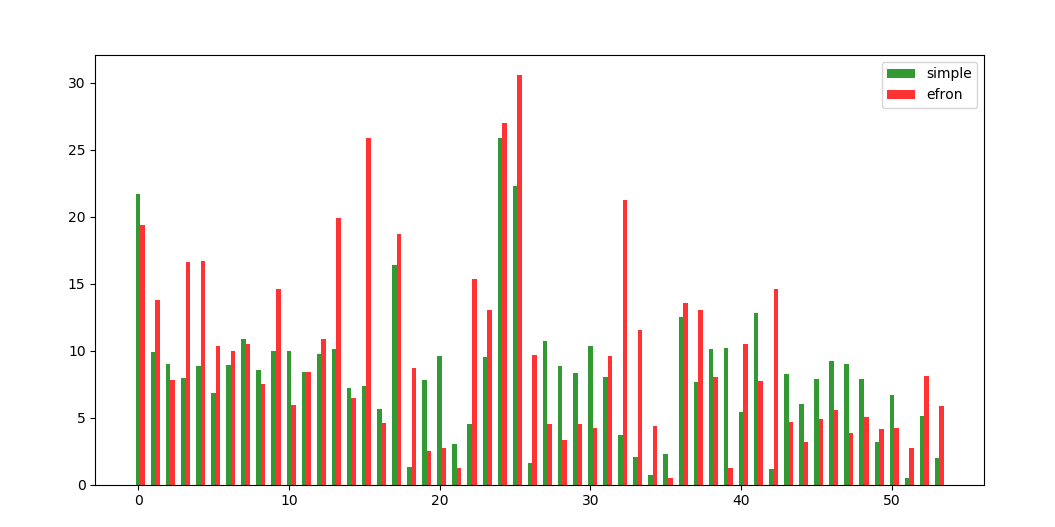


Рис. 10. Метод Виллемейна

**5.3 Сравнение результатов**

К сожалению, реализованные методы не показали выдающихся результатов. В частности, методы сглаживания, которые были призваны помочь и улучшить методы бутстрапа, не сильно помогли. Сравнение методов можно увидеть в таблице 4.

Таблица 4. Сравнение методов.

|  |  |
| --- | --- |
| Наименование метода | Процент средней ошибки среди магазинов |
| Простой метод прогнозирования | 8.2% |
| Бутстрап | 9.7% |
| Бутстрап + просто метод сглаживания | 9.6% |
| Бутстрап + Виллемейн | 9.4% |

По этим итоговым данным можно сделать, возможно, ошибочный вывод, что простой метод прогнозирования лучше, чем умные методы. Так может показаться по ряду причин:

1. Некачественная реализация с возможными ошибками.
2. Некачественные данные. Данные методы очень хорошо работают с теми продажами, где практически треть данных отсутствует. У нас же есть данные, где данные практически все имеются. Это может послужить такой большой ошибкой.
3. Так как настоящих данных нет и как сравнивать неизвестно, было решено взять последнюю неделю из обучающих данных. Возможно, в каждом из магазинов на этот период были продажи меньше или больше, чем обычно, в связи с какими-нибудь событиями в стране или ещё что-то. Что естественно будет мешать точному прогнозу. А в данных методах такие особенности не учитываются, поэтому они рассмотрены не были.

**5.4 Вывод.**

В данной главе были описаны результаты выполнения реализованных методов. Были оценены такие методы, как: простой метод прогнозирования, бутстрап. И методы сглаживания, которые должны были улучшить конечные значения при прогнозе: простой метод сглаживания, метод Виллемейна.

По итогам, метод бутстрап показал себя чуть хуже, чем метод простого прогнозирования. В половине магазинов ошибка менее 5 процентов, в половине более 10. Примерно средняя ошибка по всем магазинам 9.7 процентов. Что является довольно большим процентом ошибки.

Методы, призванные улучшить данное значение, а именно методы сглаживания, справились со своей задачей, но не сильно удачно, в лучшем случае, метод Виллемейна улучшил значение до 9.4 процентов, когда как метод простого сглаживания улучшил лишь до 9.6 процентов.

Причины таких высоких ошибок может быть множество, и это не значит, что непосредственно сами методы виноваты в такой большой ошибке.

**6. Заключение.**

В рамках дипломной работы было исследована применимость метода бутстрап и других непараметрических методов для восстановления распределения спроса на товары. Описаны и реализованы методы восстановления, произведено сравнение этих методов. Обоснованы и разработаны методы сглаживания, призванные улучшить текущие методы восстановления общего спроса продаж.

Проведённые числовые, аналитические и графические исследования, которые позволяют судить об эффективности и точности оценок, которые были получены с помощью предлагаемых подходов. Метод бутстрап и другие непараметрические методы показали относительно неплохой результат. Ошибка состояла от восьми до девяти процентов. Данные методы могли бы показать себя лучше, но такие методы показывают себя хорошо с данными, где около тридцати процентов данных являются нулевыми или же если в данных отсутствует сезонность или что-то, что с ней связано. Во взятых нами данных, к сожалению, нет провалов и так же присутствует сезонность. Возможно, именно благодаря этому получается именно эта оценка.

Кроме того, данные методы предлагают работу с различными вариантами неполноты информации о цензурировании, в то время как существуют методы, которые предлагают только наличие полной информации о цензурировании. Важно отметить, что точность оценок в случае использования неполной информации о цензурировании снижается незначительно.

Полученные результаты служат надежной основой для дальнейших исследований. Разработанные подходы к оценке общего спроса могут быть внедрены в рамках любой из функционирующих компьютерных систем управления доходами. Кроме того, ценность данных подходов состоит в том, что их можно применять при работе с цензурированными данными различной природы, не только с цензурированным спросом.

Для дальнейшего исследования в рамках задачи восстановления спроса можно отметить два ключевых направления.

Первое направление – это исключение предположения о независимости спроса на продукты различных ценовых классов. То есть подход к восстановлению спроса должен работать одновременно с несколькими связанными классами. Благодаря этому, в этот подход будет заложена возможность перехода покупателей между классами.

Второе – расширение предлагаемого подхода на случай, когда спрос представляет собой случайную величину, а пределы бронирования являются функцией от этой случайной величины.

Выведенные подходы восстановления спроса могут использовать в различных сферах рынка. От продаж еды, до продаж автомобилей и комплектующих. Но лучшим образом данные подходы показывают себя в случае редкости продаж, поэтому лучшим образом данные подходу покажут себя в индустрии изделия деталей для автомобилей, пароходов, самолётов и многого другого.

**Список литературы.**

1. Вентцель, Е.С. Исследование операций/ Е.С. Вентцель – Москва: Советское радио, 1972. – 552 с.
2. Кокс, Д. Р., Оукс Д. Анализ данных типа времени жизни/Д. Р. Кокс, Д. Оукс - Москва: Финансы и статистика. - 1988. - 192 с
3. Ибрагимов, И. Асимптотическая теория оценивания/ И. Ибрагимов, Р. Хасьминский// Москва: Наука. - 1979. - 527 с.
4. Лапина, М. Г. Построение дискретных распределений случайных величин для моделирования общего неограниченного спроса / М.Г. Лапина, Г.М. Фридман // В мире научных открытий. - 2012. - № 8 (32). - С. 210-222.
5. Лапина, М. Г. О влиянии предположения о типе распределения данных на эффективность восстановления нецензурированного спроса / М.Г. Лапина// Вестник гражданских инженеров. - 2013. - № 5 (40). - С. 288-294.
6. Belobaba, P. P. Air travel demand and airline seat inventory management: Ph.D. thesis/ Massachusetts Institute of Technology. Cambridge, Mass, USA, 1987. – 214 p.
7. Breslow, N. A large sample study of the life table and product-limit estimates under random censorship/ N. Breslow, J. Crowley// The Annals of Statistics. – 1974. – Vol.2. - №3. - P. 437–453.
8. Nezih, A., Lewis A. L. Service Parts Management /Nezih Altay, Lewis A. Litteral// Demand Forecasting and Inventory Control, USA, 1987. – 214 p.

**Приложение**

import datetime

import random

from Data import TransactionHistory

import reader

import forecasting\_methods as forecast

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

DAYS\_OF\_WEEK = 7

SHOPS\_COUNT = 54

class TransactionHistory(object):

def \_\_init\_\_(self, data):

length = len(data)

self.learning\_history = data[:length - DAYS\_OF\_WEEK]

self.last\_weak = data[length - DAYS\_OF\_WEEK:]

def get\_transactions():

return pd.read\_csv('data/transactions.csv', index\_col='date')

# Простой способ прогнозирования

def simple(data, is\_day=False):

result = sum(data[len(data) - DAYS\_OF\_WEEK:])

return result / DAYS\_OF\_WEEK if is\_day else result

# Метод Эфрона

def efron(data, is\_day=False, is\_smooth=False):

result = []

for i in range(0, ITERATION\_COUNT):

var = sum(np.random.choice(data, size=DAYS\_OF\_WEEK))

result.append(var / DAYS\_OF\_WEEK if is\_day else var)

# result = smooth(data)

# result = willemain(data)

return np.average(result)

def smooth(data):

copy\_array = np.copy(data)

copy\_array.sort()

data\_range = get\_data\_range(copy\_array)

probability = get\_probability(copy\_array, data\_range)

for key in probability:

if list(probability)[len(probability) - 1] == key: break

a, b = get\_interval(copy\_array, key)

data.append(round(a + (b - a) \* random.random()))

return data

def get\_data\_range(data):

return np.math.ceil((max(data) - min(data)) / len(data))

def get\_probability(data, data\_range):

result = {}

temp\_range = data[0] + data\_range

length = len(data)

start = 0

for i in range(length):

if data[i] > temp\_range:

result[temp\_range] = (i - start) / length

start = i

temp\_range = data\_range + data[i]

result[temp\_range] = (length - start) / length

return result

def get\_interval(data, value):

i = 0

while True:

if data[i] > value:

return data[i - 1], data[i]

i += 1

def norm(x):

return (1 / np.math.sqrt(2 \* np.math.pi)) \* (np.math.e \*\* -((x \*\* x) / 2))

def willemain(data):

for i in range(len(data) // 3):

index = random.randint(0, len(data) - 1)

x = data[index]

data[index] = x + norm(x) \* np.math.sqrt(x) + 1

return data

def get\_shops(value):

result = {}

for i in range(1, SHOPS\_COUNT + 1):

data = value.loc[i]

data.index = pd.to\_datetime(data.date)

result[i] = TransactionHistory(data)

return result

def get\_errors(shops, weak\_day=''):

simple\_error = []

efron\_error = []

for value in shops.values():

last\_weak = value.last\_weak

learning\_history = value.learning\_history

if weak\_day != '':

last\_weak = last\_weak.query('weekday==' + weak\_day)

learning\_history = learning\_history.query('weekday==' + weak\_day)

correct\_value = sum(last\_weak.transactions)

data = learning\_history.transactions

simple\_error.append(MAPE(correct\_value, forecast.simple(data, weak\_day != '')))

efron\_error.append(MAPE(correct\_value, forecast.efron(data, weak\_day != '', True)))

return simple\_error, efron\_error

def MAPE(correct\_value, forecast\_value):

return abs(correct\_value - forecast\_value) / correct\_value \* 100

def show\_histogram(data):

data.plot()

plt.show()

def show\_bar(simple\_error, efron\_error):

opacity = 0.8

width = 0.3

plt.bar(np.arange(len(simple\_error)), simple\_error, width, alpha=opacity, color='g', label='simple')

plt.bar(np.arange(len(efron\_error)) + width, efron\_error, width, alpha=opacity, color='r', label='efron')

plt.legend()

plt.show()

if \_\_name\_\_ == '\_\_main\_\_':

transactions = reader.get\_transactions()

transactions.sort\_index()

transactions = transactions.loc['2013-01-07':'2017-08-13']

transactions = transactions.reset\_index()

transactions.index = transactions.store\_nbr

transactions['weekday'] = [datetime.datetime.strptime(x, '%Y-%m-%d').weekday() for x in transactions.date.values]

shops = get\_shops(transactions)

simple\_error, efron\_error = get\_errors(shops)

show\_bar(simple\_error, efron\_error)