Федеральное государственное бюджетное образовательное учреждение высшего образования Удмуртский государственный университет

На правах рукописи

Пивкин Кирилл Сергеевич

МОДЕЛИРОВАНИЕ ПОКУПАТЕЛЬСКОГО СПРОСА НА ПРЕДПРИЯТИЯХ РОЗНИЧНОЙ ТОРГОВЛИ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

08.00.13 — Математические и инструментальные методы экономики

ДИССЕРТАЦИЯ

на соискание учёной степени

кандидата экономических наук

Научный руководитель:

д. ф.-м. н., профессор

Лётчиков Андрей Владимирович

Ижевск 2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность исследования. Конъюнктура розничной торговли является ключевым индикатором развития экономики страны. Как одна из самых подвижных сфер экономики розничная сфера ярко отражает работу экономических законов как микро-, так и макроуровня. После испытаний мирового финансового кризиса и современного структурно-экономического кризиса в России конкурентоспособность отечественных предприятий розничной торговли стала, прежде всего, определяться способностью оптимизировать внутренние бизнес-процессы и потоки товародвижения. Это приводит к созданию новых методов и подходов в политике организации розничного бизнеса, в частности, в области планирования, прогнозирования и организации товародвижения.

Стохастичность спроса накладывает большие ограничения на работу розничной компании. При отсутствии системы прогнозирования должного качества страдает большинство областей управления розничного предприятия, например, область управления запасами, где низкая точность решения может привести к снижению конкурентоспособности. При этом, для торговых компаний с узким ассортиментом наиболее популярных брендов подобная проблема менее актуальна – она решается с применением дешевого труда менеджеров по ежедневному контролю низко вариативного спроса. С увеличением количества товарных позиций, увеличением торговых площадей, точек продаж и масштабов торговли обеспечить качественное прогнозирование спроса путем прямых калькуляций невозможно на практике. Поэтому важным является совершенствование текущих технологий прогнозирования и планирования покупательского спроса, автоматизации этих бизнес-процессов, причем на основе продвинутого статистического моделирования.

Современные методы прикладной статистики получили обширное применение во многих отраслях экономики ввиду технологического бума. Развитие скорости и эффективности вычислительных алгоритмов на сегодняшний день позволяет обрабатывать большие массивы данных (Big Data) даже на персональных компьютерах с помощью открытого программного обеспечения. Имея основания для разработки эффективной платформы моделирования покупательского спроса, ставится задача о повышении точности прогнозирования. Исходя из наличия проработанных в теории и на практике методов – специальных методов линейной регрессии, деревьев решений и их композиций, метода опорных векторов, нейронных сетей, градиентного бустинга и т.п. – задача высокоточного моделирования потребности представляется достижимой. Наиболее качественный прогноз позволяет решить проблему стохастичности процесса товародвижения. Затем результат моделирования предприятие может

использовать в процессах управления запасами, в планировании маркетинговых мероприятий, в динамичном управлении ценами и других областях управления.

Для построения развитой системы прогнозирования покупательского спроса ощущается острая потребность в реализации модели прогнозирования на основании методов современной теории эконометрики, статистического и машинного обучения, а также элементов теории экономики торговли и управления запасами, при этом имея конкретную форму в виде реализованного программно-вычислительного комплекса, интегрированного в информационные системы предприятия. Наличием данной потребности и обусловлена актуальность настоящего диссертационного исследования.

Степень разработанности проблемы определена как большим количеством литературных источников и публикаций по экономике торговых организаций, эконометрике и теории машинного обучения, так и достаточным количеством проработанных концепций для математического прогнозирования и моделирования, большая часть которых не универсальна и сложна во внедрении для решения бизнес-задач.

Основные понятия экономики торговых организаций даны в трудах отечественных авторов и авторов ближнего зарубежья: Р. П. Валевича [13], Г. А. Давыдовой [13], М. С. Абрютиной [1], Л. А. Брагина [85], Т. П. Данько [85], Л. А. Козерод [39], Т. М. Безбородова [9], М. Б. Дюжева [9], В. В. Лукинского [48]. В данных источниках раскрыты принципы работы торговых организаций, в частности компаний розничной торговли, а также описана ключевая роль управления товарными запасами в системе менеджмента торговой организации. В трудах зарубежных авторов касательно розничной торговли отражена информация о новых технологиях организации бизнеса и управления товарными запасами: Э. Голдратт [19], Т. Уоллас [86], Р. Сталь [86], F. Caro [101], J. Gallien [101], C. Crum [103], G. Palmatier [103], N. Lichtenstein [114], S. Tayur [123], R. Ganeshan [123], M. Magazine [123], D. Bartmann [99], M. F. Bach [99]. Тем не менее, во многих из этих источников уделяется недостаточно много внимания важности прогнозирования спроса в розничном бизнесе, либо в его методологическом описании в качестве инструментов используются упрощенные экономико-математические модели, не учитывающие возможные нелинейные связи между факторами.

Прогнозирование социально-экономических и иных показателей, которые можно охарактеризовать таким общим понятием как временной ряд, описано в исследованиях ряда зарубежных и отечественных ученых: Р. Хиндмана [117, 109], Дж. Атанасопулоса [109], J. S. Racine [110, 111], Q. Li [113], T. Hayfield [110], С. А. Айвазяна [2, 3], Б. Б. Демешева [12], И. С. Светунькова [75, 76, 120, 121], С. Г. Светунькова [75, 76], N. Kourentzes [120], А. И. Орлова [58], В. Н. Афанасьева [5], М. М. Юзбашева [5], О. А. Мишулиной [53], О. М. Писаревой [71], Н. А. Садовниковой [74], Р. А. Шмойловой [74], Т. А. Дубровой [31], В. К. Семёнычева [78], Е. В.

Семёнычева [78]. Исследования посвящены изучению динамики социально-экономических показателей, их свойствам и математическому моделированию процессов и временных рядов. Представлен широкий обзор тренд-сезонных и авторегрессионных инструментов, но практически не рассматривается сложное многоуровневое моделирование динамических процессов.

Темы продвинутого математического прогнозирования, статистического и машинного обучения, которые тесно связаны с задачами прогнозирования и оптимизации в розничной торговле проработаны в ведущих исследованиях современных зарубежных и отечественных авторов: Г. Джеймса [27], Д. Уиттона [27], Т. Хасти [27, 125], Р. Тибширани [27], Э. Ына [115], Л. Бреймана [100, 104], А. Мюллера [54], С. Гвидо [54], У. Маккинли [51], С. Рашка [72], Я. Гудфеллоу [22], И. Бенджио [22], А. Курвилля [22], Дж. Хинтона [118], П. Флаха [89], Д. Кука [43], Л. П. Коэльо [41], П. Домингоса [28, 106], С. Осовского [59], P. Cichosz [102], Ю. И. Журавлева [33, 57], Л. Н. Ясницкого [98], К. В. Воронцова [17, 18], Д. П. Ветрова [15], А. В. Груздева [21], А. Б. Меркова [52]. В источниках раскрыты основные достижения в области компьютерно-математического моделирования процессов и явлений тех процессов, которые описываются достаточно большим массивом учетных данных. Представленные примеры задач из практики ограничены классическими примерами статистического и машинного обучения: биологические задачи, задачи распознавания изображений, прогнозирование несложных процессов, т.е. отсутствует глубокое погружение в ту или иную область применения, в которой может быть использовано продвинутое математическое прогнозирование.

На текущий момент на отечественном рынке реализован ряд программных продуктов по прогнозированию спроса и управлению запасами такие как «Forecsys Goods4Cast», «Forecast NOW!», «Deductor», «Прогноз». На основе разработок и решений, реализованных с помощью рассмотренных продуктов, написано ряд статей таких авторов как Н. Б. Паклина [60, 83], В. И Орешкова [60], Ш. Акобира [4], Бариновой О. В. [7], А. А. Грицая [20] и многих других. Источники посвящены в основном методикам исследования данных, прогнозированию спроса и математическим алгоритмам, лежащим в основе программных комплексов по прогнозированию спроса. Большинство из рассмотренных программных приложений разработаны по концепции «черного ящика», когда пользователь не может влиять на структуру разработанных алгоритмов.

Проблема разработки информативных моделей, апробации и внедрения программных решений в области прогнозирования спроса на предприятии розничной торговли определило выбор объекта, предмета, цели и задач диссертационного исследования.

Объектом исследования является предприятие розничной торговли как система товарных потоков и информации о них.

Предметом исследования является процесс товародвижения предприятия розничной торговли.

Целью диссертационного исследования является теоретическое и практическое развитие прогнозного моделирования покупательского спроса на предприятиях розничной торговли на основе методов машинного обучения.

Для достижения цели в данной работе необходимо решить следующие исследовательские и практико-ориентированные **задачи**:

1. Построить эмпирическую модель прогнозирования товарного спроса на основе данных пространственно-временной выборки.
2. Построить систему прогнозирования для переменных, которые включены как предикторы в основную модель спроса розничного предприятия.
3. Разработать программный комплекс для оценки будущего розничного спроса на основе построенной модели прогнозирования.

Область исследования соответствует паспорту научной специальности ВАК РФ 08.00.13 «Математические и инструментальные методы экономики» по следующим пунктам:

1.4. Разработка и исследование моделей и математических методов анализа микроэкономических процессов и систем: отраслей народного хозяйства, фирм и предприятий, домашних хозяйств, рынков, механизмов формирования спроса и потребления, способов количественной оценки предпринимательских рисков и обоснования инвестиционных решений.

2.3. Разработка систем поддержки принятия решений для рационализации организационных структур и оптимизации управления экономикой на всех уровнях.

Теоретическая и методологическая основа исследования

Теоретическую основу данной диссертационной работы составляют исследования в области торгового дела и экономико-математического прогнозирования. Для структурного анализа объекта исследования применяется инструментарий экономики торговых организаций с целью обозначить принципы функционирования торгового предприятия и область применения разрабатываемой экономико-математической модели. Используемый инструментарий для решения основной задачи – создания модели прогнозирования спроса – основан на классических и современных методах машинного обучения: классической линейной регрессии, линейной регрессии с регуляризацией, деревьях решений и их ансамблевых реализаций в виде случайного леса и бустинга.

Научная новизна. В ходе проведенного исследования получены результаты, которые обладают научной новизной и являются предметом защиты:

1. Построена модель прогнозирования спроса конкретной товарной позиции на основе пространственно-временной выборки данных с применением современных методов

машинного обучения, позволяющая учитывать, в отличие от существующих регрессионных моделей прогнозирования спроса, особенности мультимодального спроса на товар (1.4. Разработка и исследование моделей и математических методов анализа микроэкономических процессов и систем: отраслей народного хозяйства, фирм и предприятий, домашних хозяйств, рынков, механизмов формирования спроса и потребления, способов количественной оценки предпринимательских рисков и обоснования инвестиционных решений. Глава 2, параграфы 2.5 и 2.6, стр. 61-70).

2. Построена оригинальная методика расчета будущих значений ключевых переменных математической модели прогнозирования спроса, увеличивающая прогностическую точность за счет использования современных инструментов анализа временных рядов: метод Prophet, байесовские временные ряды, современные версии алгоритмов ARIMA и экспоненциального сглаживания. (1.4. Разработка и исследование моделей и математических методов анализа микроэкономических процессов и систем: отраслей народного хозяйства, фирм и предприятий, домашних хозяйств, рынков, механизмов формирования спроса и потребления, способов количественной оценки предпринимательских рисков и обоснования инвестиционных решений. Глава 2, параграф 2.4, стр. 52-61).
3. Разработан программный комплекс прогнозирования спроса на основе языка программирования R, функционирующий как сервис, встроенный в автоматическую систему заказа товара на предприятии розничной торговли, и способствующий удовлетворению покупательского спроса и оптимизации товарных запасов (2.3. Разработка систем поддержки принятия решений для рационализации организационных структур и оптимизации управления экономикой на всех уровнях. Глава 3, параграфы 3.4, 3.5, стр. 111-116).

Научная и практическая значимость результатов исследования

Научная значимость исследования состоит в критической оценке текущих методов и моделей прогнозирования товарного спроса и разработке новой экономико-математической модели, которая позволяет учитывать динамику спроса и экзогенных факторов. Здесь учитывается эффект многономенклатурности выбора товаров на итоговую потребность покупателя и эффекты, которые связаны с покупательским поведением, эластичностью спроса. Результаты, полученные в работе, вносят вклад в решение одной из самых важных народно-хозяйственных проблем повышения эффективности инструментов прогнозирования спроса и оптимизации управления товарными запасами в качестве приложения результатов моделирования. Практическая значимость работы заключается в реализации программного продукта в виде сервиса на языке R, который в автоматическом режиме осуществляет

прогнозирования потребности покупателей в товаре. Алгоритм в основе работы сервиса является универсальным для торговых сетей, продающих товары повседневного спроса.

Апробация результатов исследования. Результаты диссертационного исследования были представлены на Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2017» (Москва, 2017 г.), Международной научно-технической конференции «Перспективные информационные технологии» (Самара, 2017 г.), Всероссийской заочной научно-практической конференции «Математические методы и интеллектуальные системы в экономике и образовании» (Ижевск, 2015 г., 2016 г., 2017 г.), Международной молодёжной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками» (Саратов, 2017 г.). Результаты исследования использовались при проведении занятий по дисциплинам «Моделирование бизнес-процессов», «Эконометрическое моделирование» и «Информационные системы управления производственной компанией» для студентов бакалавриата по направлению «Бизнес-информатика» Института экономики и управления Удмуртского государственного университета.

Результаты диссертационного исследования находятся на стадии активного внедрения в бизнес-процессы розничного предприятия ООО «Гастроном». Применяя разработанные в диссертационной работе экономико-математические модели, ритейлер улучшил качество прогнозирования и разработки планов на ключевые показатели предприятия, увеличил уровень продаж по ряду товарных групп и оптимизировал товарный запас, что подтверждено Актом о внедрении результатов диссертационной работы.

Публикации. По теме диссертационного исследования опубликовано 10 работ объемом 7,11 п. л., из них в ведущих рецензируемых научных журналах и изданиях, определенных ВАК, - 4.

Структура и объем работы. Диссертация содержит введение, 3 главы и заключение, изложенные на 145 страницах машинописного текста. В работу включены 57 рисунков, 20 таблиц, 6 приложений и список литературы из 125 наименований.

Введение содержит описание актуальности темы, формулировку целей и задач работы, раскрывает основные научные методы, используемые в работе, основные положения, выносимые на защиту, и определяет содержательную часть работы.

В **первой главе** проведен анализ системы товародвижения на предприятии розничной торговли. Рассмотрено понятие покупательского спроса, его ключевые аспекты, а также важность задачи моделирования и прогнозирования. Сделан акцент на особенностях моделирования покупательского спроса в рамках настоящего диссертационного исследования. Приведены математические методы, которые рассматриваются как основные инструменты моделирования и прогнозирования спроса.

Во **второй главе** разработана методология прогнозирования ключевых переменных и прогнозирования целевой переменной – покупательского спроса. Модели прогнозирования ключевых показателей выстроены на классических и современных методах прогнозирования временных рядов. Разработанная модель прогнозирования спроса базируется на работе с панельными выборками, кластеризации временных рядов, обогащении исходных данных переменными экономического толка и применении методов машинного обучения. Приведенные модели оценены с помощью стандартных метрик качества для регрессий.

В **третьей главе** на основании данных, предоставленных ООО «Гастроном», рассчитаны оценки параметров и гиперпараметров приведенных регрессионных моделей. Оценена совмещенная модель из разных представленных методов на основе средневзвешенных оценок прогноза. Получены результаты по прогнозированию спроса и проведен сравнительный анализ с фактической ситуацией, на основе которого делаются выводы об экономической эффективности проведенной работы. Отражена архитектура и программная среда разработанной системы поддержки принятия решения, созданной на основе описанной методологии.

В **заключении** содержится описание основных выводов и результатов исследования.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ МОДЕЛИРОВАНИЯ ПРОЦЕССА ТОВАРОДВИЖЕНИЯ

1.1. Характеристика товарно-розничного предприятия как открытой системы товародвижения

На данный момент совершенствование систем, способствующих улучшению эффективности потоков товародвижения на предприятии розничной торговли, является необходимым элементом стратегии ведения бизнеса. В условиях современной экономической ситуации: волатильности курсов валют, структурных изменений потребительского спроса, продовольственного эмбарго и многих других конъюнктурных проблем, эффективные методы, используемые в области моделирования и прогнозирования товарных потоков, дают розничным сетям необходимое конкурентное преимущество. Очевидно, что в подобных условиях инструменты математического моделирования позволяют связать многие внутренние и внешние факторы, влияющие на работу в розничной организации, в единую систему планирования и поддержки принятия решений.

Поэтому, для того чтобы приступить к детальному исследованию проблем в построении системы прогнозирования покупательского спроса, необходимо определить основные аспекты работы торгово-розничного предприятия. Для начала необходимо раскрыть природу понятия торговли как таковой. В большинстве источников [9, 39] торговля определяется как крупный сектор экономики, который выполняет функцию по купле-продаже (обмену) товарно-материальных ценностей среди субъектов экономики. Учитывая историческую и социально-экономическую значимость данного сектора, его состояние влияет на уровень жизни населения, развитие экономики, технологий и многие другие аспекты современности. При этом в указанной системе, торговое предприятие выступает в роли агента – посредника между производителем блага и его промежуточным или конечным потребителем. Цель торгового предприятия – получить прибыль на операциях купли-продажи, при этом оптимизируя издержки для повышения своей конкурентоспособности.

Основная особенность розничной торговли, в отличие, например, от оптовой, определяется в том, что продаваемый товар ориентирован на конечного потребителя. Соответственно, это накладывает отпечаток на деятельность розничных предприятий – в своих приоритетах они ориентируются прежде всего на развитие маркетинговой составляющей в стратегии бизнеса; систематизируют подход к конечному покупателю и имеют более динамичную структуру продаж во времени. Исходя из анализа источников [11, 61] необходимо выделить несколько задач розничного предприятия:

- изучение и сбор информации о конечном потребителе товара;
- организация отношений с иными участниками сферы товарного обращения, в первую очередь – с поставщиками товара;
- заказ товара и организация его хранения на специализированных складских помещениях;
- организация торгового пространства для выкладки имеющихся благ для взаимодействия с конечным потребителем (покупателем);
- оптимизация внутренних бизнес-процессов предприятия;
- разработка и осуществление стратегических и маркетинговых целей розничного предприятия.

Видно, что спектр задач разнообразен и требует значительных усилий от менеджмента торгово-розничного предприятия. Ситуация осложняется тем, что большинство торговых предприятий имеет узкую специализацию, которая зависит от ассортимента продаваемых товаров. Под ассортиментом понимается довольно разнообразный набор благ с разным назначением для потребителя, который торговый посредник приобретает с целью дальнейшей продажи. По Памбухчиянцу [61] можно выделить четыре группы в специализации розничных предприятий:

- продовольственные;
- непродовольственные;
- смешанные;
- розничные предприятия прочей специализации.

В большей доле на розничном рынке оборачиваются товары повседневного пользования, т.е. речь идет о так называемом рынке FMCG (от англ. fast moving consumer goods – быстро оборачиваемые потребительские товары). Здесь, например, ведется обмен продуктов питания, предметов личной гигиены, моющих средств, косметики и других товаров с повседневной потребностью. Но также стоит отметить отдельный класс товаров длительного пользования, куда входят бытовая техника, автомобили и другие, связанные со среднесрочным и инвестиционным потреблением общества. Группа таких товаров характеризуется высокими вложениями потребителя продукта и низкой частотой покупки (обычно, благо длительного пользования обменивается не чаще одного раза в год).

Исходя из понятия о розничной торговле и о розничном предприятии, можно представить последнее как поток товарно-материальных ценностей, который переходит от поставщика товара к конечному потребителю, путем последовательного исполнения задач предприятия. Существующий поток товарно-материальных ценностей, находящийся в обращении, формирует уровень товарного запаса у конкретного розничного предприятия.

Для понимания определений и сути товародвижения на предприятии розничной торговли необходимо рассмотреть упрощенную схему движения товарно-материальных ценностей для розничной торговли [65, 66], представленную на рис. 1.1.

Рисунок 1.1 – Движение товарно-материальных ценностей в системе розничной торговли

Несложно определить, что интенсивность и характер движения товаров в данной цепочке (при условиях рыночной экономики) задает конечный потребитель – покупатель товаров в розничном магазине. Соответственно, розничное предприятие формирует свой ассортимент, осуществляет маркетинговую стратегию, организует промо-активность, реализует закупочные программы и разовые заказы товаров у поставщика, прежде всего отталкиваясь от уровня и качества покупательского спроса. Кроме того, это вполне логично и с точки зрения оптимизации затрат. Например, в области формирования товарных запасов на основании некорректной оценки спроса на собственную продукцию розничная сеть:

- либо формирует запасы значительно выше уровня потребления, что приводит к дополнительным затратам на хранение и к порче товара (что особенно актуально для товаров с низкими сроками реализации) или потери его востребованности;
- либо формирует запасы ниже уровня потребления, что приводит к так называемому *out-of-stock*, т.е. отсутствию товарного запаса на складе и на витринах магазина. Как следствие – упущенные доходы торгового предприятия, а также имиджевые издержки, так как бизнес розничного продавца подразумевает под собой, прежде всего, гарантию наличия товара, пользующегося спросом.

Следует также отметить, что розничная компания чаще всего имеет непростую внутреннюю и внешнюю структуру потока товародвижения. В формате розничного предприятия сложной управленческой и операционной структурой обладает торгово-розничная сеть. Розничная сеть определяется как совокупность нескольких торговых точек (магазинов) объединенных территориальным признаком, которые имеют единую стратегию управления, и цель которой состоит в получении прибыли с помощью операций купли-продажи [85]. Ключевым моментом здесь является понятие совокупности магазинов, которое ставит перед управлением розничным предприятием новые вопросы: управления персоналом, неоднозначной эффективности маркетинговых мероприятий для разных элементов сети, качественного взаимодействия с поставщиками товаров, взаимодействия магазинов между собой и центральным аппаратом управления организацией. Ниже приведены схемы возможных структур

розничной сети с точки зрения степени централизации движения товарно-материальных ценностей внутри сети.

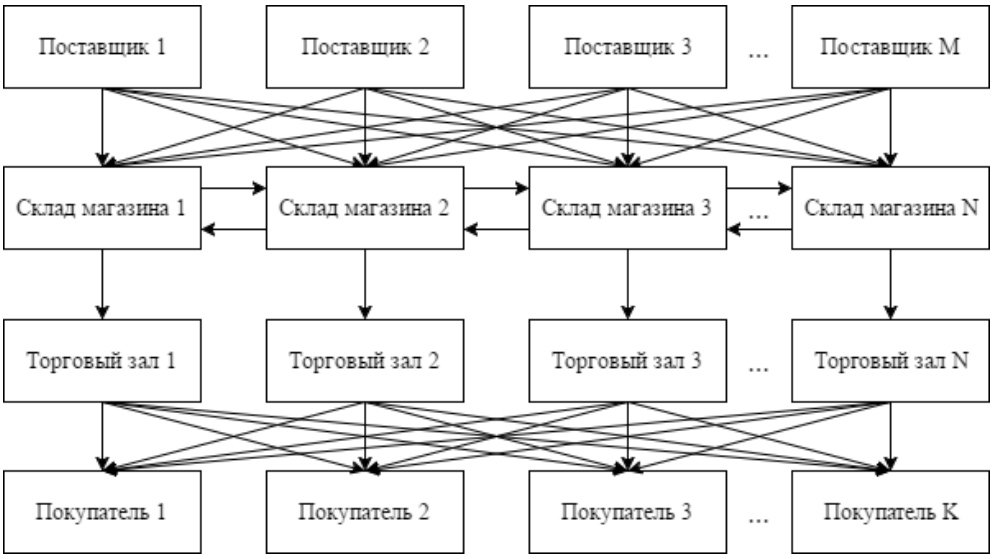


Рисунок 1.2 – Децентрализованный поток товародвижения

Исходя из схемы децентрализованного потока товародвижения, представленной на рис. 1.2, товарно-материальные ценности движутся от M поставщиков к N складам магазинов розничной сети, избегая посреднических операций. Подобная структура требует четко согласованной работы с поставщиками товаров для минимизации рисков непоставки товара и *out-of-stock*. На рисунке 1.2 также изображено движение товаров между складами магазина (для упрощения выведен циклически, но обычно взаимодействие происходит между всеми складами). При наличии условия обмена товарными запасами между складами магазинов появляется компенсационный эффект в случае неправильного заказа товара и/или непоставки товара со стороны поставщика в срок. Таким образом, магазин-заказчик может получить недостающий товарный запас от магазина-поставщика, тем самым продолжая получать доход от продаж. Тем не менее, в такой ситуации возрастают затраты на транспортировку товара между магазинами сети.

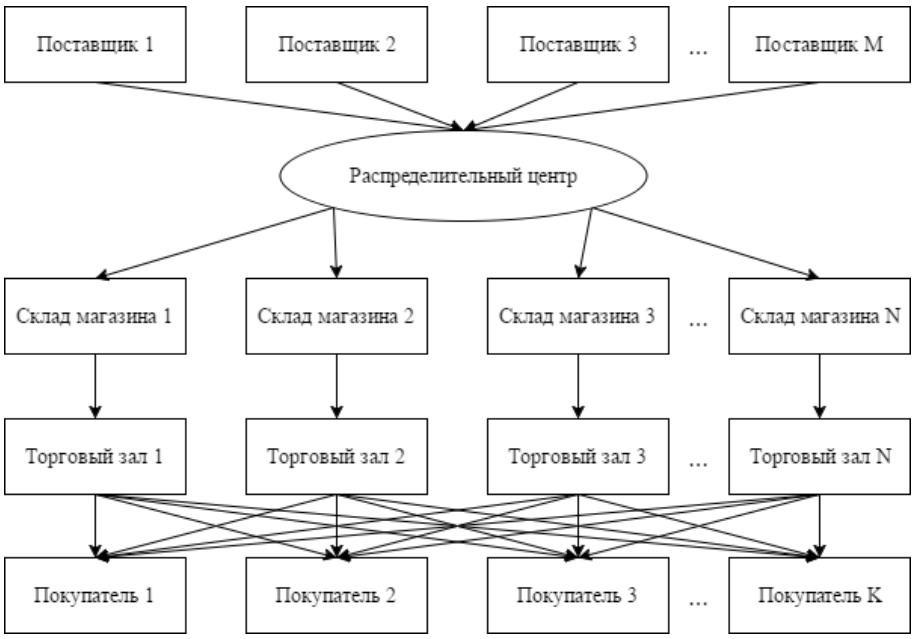


Рисунок 1.3 – Централизованный поток товародвижения

Централизованный поток товародвижения, который виден на рис. 1.3, отличается от децентрализованного наличием важного элемента – *распределительного центра* розничной сети. По Леви и Вейтцу [46] распределительным центром называется склад розничного предприятия, на который осуществляется доставка товара от нескольких поставщиков и с которого, затем, отправляется запас для складов розничной сети. Наличие распределительного центра вносит существенный вклад в минимизацию издержек по транспортировке товара, снижает управленческие риски и риски нехватки товара.

В рамках настоящего диссертационного исследования типы товародвижения внутри розничной сети отражают высокую сложность товарных потоков. Информация на рис. 1.2 и рис. 1.3 помогает понять, что розничная сеть имеет N торговых залов и K покупателей, которые также постоянно мигрируют от одной торговой точки до другой. Это означает, что покупательский спрос дифференцирован как по времени, так и по пространству: одна и та же ассортиментная матрица имеет разный спрос в рамках разных торговых залов в разные моменты времени. И если в случае централизованной схемы стоимость ошибки в прогнозах потребности покупателя не так высока, то в децентрализованных торговых сетях при ошибочных оценках спроса стоимость дополнительной единицы товара для удовлетворения спроса свыше прогноза может быть очень высокой. Это дает представление о том, что задачу прогнозирования спроса необходимо решать, учитывая структуру торгово-розничного предприятия и особенности товародвижения внутри.

Для того чтобы подытожить основное представление о розничном предприятии как системе товародвижения, необходимо также обозначить его ключевые показатели. В теории и практике розничной торговли главным показателем эффективности работы предприятия является товарооборот – суммарный объем продаж товаров конечному потребителю

(покупателю) в денежном выражении за указанный период времени [34, 95]. Розничный товарооборот определяет объемы продаваемых товаров, выражает весь товарный поток, который существует на предприятии в денежных единицах. Это позволяет производить экономическую оценку управленческих решений, соотнося динамику товарооборота, текущие постоянные и переменные издержки предприятия. Еще одним важным показателем является уровень торговой надбавки (наценки) к закупочной стоимости товара. Торговая наценка применяется к цене товара для покрытия расходов розничного предприятия и получения прибыли. Существует множество методик определения торговой надбавки, связанных в первую очередь с затратными и рыночными аспектами розничного бизнеса. Торговая наценка является целевым показателем для розничной торговли, так как она определяет валовый доход розничной компании по реализации продукции [49]. Рассмотренные показатели очень важны для самоопределения розничного продавца как организации отличной от иных сфер экономической деятельности. С помощью показателей товарооборота и наценки также можно вывести приоритетные направления в оптимизации товарных потоков на предприятии, что напрямую связано с задачей управления запасами.

Характеристика предприятия розничной торговли как открытой системы товародвижения со своими значимыми структурными элементами – поставщик, склад, торговый зал, покупатель – позволила обозначить условия, в которых формируется покупательский спрос на тот или иной товар. По сути, если сами товарные потоки идут от поставщика к покупателю (см. рис. 1.1, 1.2 и 1.3), то важная информация о величине, качестве и интенсивности товарных потоков подается конечными потребителями. Для розничной сети основной задачей является стремление «угадать желания» покупателя, его предпочтения для того, чтобы обеспечить стабилизацию товарных потоков внутри системы. Формально, речь идет о создании системы прогнозирования спроса, результаты которой обеспечивают корректность действий в сферах маркетингового продвижения товара, управлении запасами, системе планирования на розничном предприятии. Для полного представления о целях диссертационной работы необходимо обозначить понятие и особенности покупательского спроса, а также смысл моделирования и построения прогноза.

1.2. Моделирование покупательского спроса как основная задача автоматизации процессов розничного магазина

В условиях рыночной экономики одной из важных задач коммерческих организаций является удовлетворение потребительского спроса на продукты и услуги, которые они производят. При этом, под потребительским спросом обычно понимается «запрос фактического или потенциального потребителя на приобретение товара за имеющиеся у него, предназначенные

для покупки этого товара деньги» [11]. Потребительский спрос, его структура и интенсивность определяет конъюнктуру экономики в целом. В частности, он является базовым для показателя совокупного спроса и для расчета валового внутреннего продукта как одной из основных метрик экономического состояния.

Многие исследования в области потребительского спроса, посвящены определению факторов, детерминант, условий, которые формируют потребительский спрос. Например, Денисов обозначает следующие условия, детерминирующие потребительский спрос [25, 26]: экономические условия, политические условия, социально-институциональные условия, геологические условия и ситуационные условия. Разнообразие возможных условий говорит о сложности процессов, которые определяют потребительский спрос. При этом фундаментальность понятия в экономике, позволяет считать его одним из наиболее значимых в изучении. По итогам строятся модели потребительского спроса, которые учитывают внешние условия, и позволяют использовать результат для развития экономической политики и стратегии на мезо- и макроуровнях, обычно в виде создания прогнозов ситуации и реализуемых планов на их основе.

В рамках данной диссертационной работы рассматривается похожий подход, который распространяется на решение проблемы моделирования спроса для предприятий розничной торговли. Для этого уточняется терминология спроса: вместо потребительского рассматривается покупательский спрос. Здесь определяющим отличием является то, что конечным потребителем товара является именно покупатель розничного магазина. Кроме того, на уровне розничного предприятия важно рассматривать покупательский спрос в связке с конкретным товаром или товарной группой, так как это определяется характером товарных потоков (рис. 1.1). То есть, предлагая на своих торговых площадях новый продукт, розничному торговцу важно знать ответ на вопрос «сколько штук товара N продастся за период T?». В такой постановке покупательский спрос выражен в потребительском спросе группы лиц, которые желают и имеют возможность приобрести указанный товар. Высокая точность ответа на этот вопрос позволит принять более эффективные решения о вводе продукта, о формировании заказа на него у поставщика и определить возможные стратегии маркетингового продвижения.

Покупательский спрос в рамках розничного магазина зависит от многих факторов:

1. Бренд-менеджмент магазина, его восприятие населением. Здесь имеет значение образ торговой компании у потенциального или реального покупателя.
2. Бренд-менеджмент товара, восприятие его качества. Речь идет о торговой марке продаваемого товара, его ценности и качестве для конкретного потребителя.

3. Маркетинговая политика торгово-розничного предприятия. В первую очередь рассматривается влияние ценовых факторов на покупательский спрос, а также промо-активности розничной компании и компаний конкурентов.
4. Реализация мероприятий по мерчандайзингу товара. Корректная подача товара покупателю, выбор места продажи и правильная визуализация товара и торговых площадей.
5. Финансово-экономические отношения между розничным торговцем и его поставщиками. Определенные условия работы с поставщиком, которые влияют на характер товародвижения в компании: ограничения по доставке, скидки, штрафы, закупочные программы.
6. Организация работы персонала. Наличие активного продвижения товара в магазине
7. Географическое положение магазина. Расположение магазина в активных зонах покупательского потока.

Перечисленные факторы являются специфическими для розничного продавца, но не являются исчерпывающими. Как было оговорено ранее, возможны какие-либо общие или ситуативные условия, например, влияние политических решений на покупательский спрос в целом.

Методы изучения спроса также достаточно разнообразны [16]:

- Анализ статистической отчетности.
- Статистические данные о доходах и бюджетах домохозяйств.
- Анкетные опросы, интервью, специальные эксперименты и наблюдения.
- Методы, которые сочетают в себе ряд перечисленных.

Однако, если статистические данные являются низкозатратным источником информации о спросе, то методы, основанные на прямом контакте с покупателем, являются довольно дорогими и редкими.

Задачи компании розничной торговли, описанные в разделе 1.1, позволяют обозначить, что изучение покупательского спроса необходимо для ежедневного принятия решений. Если на макроуровне учет статистики показателей, характеризующих потребительский спрос, происходит за более длительные периоды, то на уровне частного бизнеса учет проводится ежедневно или чаще. Это нужно для своевременной оценки изменений в покупательском поведении. Соответственно, в рамках данной работы следует уточнить, что моделирование покупательского спроса и последующее построение прогнозов интересно для краткосрочного горизонта: это нужно для ежедневной корректировки принятых решений розничными компаниями в соответствие с изменениями в прогнозе.

Кроме того, современные конкурентные условия требуют от розничной торговли более высокой степени автоматизации бизнес-процессов. Ввиду этого, процесс построения прогнозов покупательского спроса последовательно переходит в область автоматизированного построения математических моделей и их использования в виде сервисов ПО. Становится актуальным использование открытых источников информации, машиночитаемой информации при моделировании покупательского спроса.

В контексте улучшения процессов товародвижения моделирование покупательского спроса воспринимается не столько как маркетинговая задача, которая позволяет выделить основные детерминанты спроса на товар, а сколько как задача прогнозирования и автоматизации процессов поддержки принятия решения. Это особенно актуально для краткосрочных горизонтов прогнозирования, которые охватывают небольшой период постановки бизнес-целей (до одного месяца). Реализация задачи моделирования в виде сервиса накладывает ограничения на входящие переменные модели – могут быть использованы в основном статистические и учетные данные, которые формируются на ежедневной основе. В таком случае, основным требованием к моделированию является получение более точного результата с точки зрения прогноза покупательского спроса. Подобный результат может быть достигнут только с помощью применения продвинутого математического и статистического инструментария, который реализуется в виде программного кода. На данный момент это является основой машинного обучения и всех программных сервисов, которые базируются на нем. Применение этой связки инструментов позволит создать сервис прогнозирования покупательского спроса, который будет унифицированным в использовании для многих участников отрасли розничной торговли. Это является фундаментом для развития технологий в розничной торговле и экономике в целом.

В следующем разделе определяются составляющие понятийного аппарата математического прогнозирования спроса в розничной торговле, а также приводятся классические методы его прогнозирования.

1.3. Основы математического прогнозирования спроса на товар

Для однозначного определения понятия прогнозирования спроса на товар необходимо отразить общее понятие прогнозирования в современной научной среде. В энциклопедических источниках прогнозирование характеризуют как «определение тенденций и перспектив развития тех или иных процессов на основе анализа данных об их прошлом и нынешнем состоянии» [96]. В целом, это является достаточно емким и цельным определением процесса прогнозирования, отражающим его основную суть.

Еще одним емким определением, сформулированным в форме вопроса, является определение из Энциклопедии эпистемологии и философии науки [97]: прогнозирование – это «рассуждение, направляемое вопросом: «Какое событие (явление, ситуация, свойство, процесс) возможно, если имеет место ситуация q?». Здесь важным моментом является, то что ситуация q определена набором характеристик, известных исследователю. Соответственно, для того чтобы понять вероятность наступления события, необходимо формализовать его *модель*, куда включена ситуация q в связке с указанными характеристиками.

В экономико-математической терминологии (Лопатников, [47]) прогнозирование – это «система научных исследований качественного и количественного характера, направленных на выяснение тенденций развития народного хозяйства или его частей (отраслей, регионов, предприятий и т.п.) и поиск оптимальных путей достижения целей этого развития». Подобное понятие отражает конкретные аспекты прогнозирования – наличие качественных и количественных выводов, на основании которых осуществляется прогноз, но при этом включает в себя также процедуру управления – в этом случае поиск способов достижения значений полученного прогноза при осуществлении задач. В указанной трактовке понятие прогнозирования становится довольно близко к понятию планирования.

Современное понятие прогнозирования определено Светуньковым И. Г. и Светуньковым С. Г., как «результат именно индуктивного вывода, когда по характеру ограниченного множества значений показателей или взаимосвязи факторов делается вывод о том, что и остальные, еще не наблюдаемые значения этих показателей или взаимосвязи будут обладать аналогичными свойствами» [76]. Как и в предыдущих определениях, понятие подразумевает под собой построение модели, которая характеризует исследуемый объект, на основе имеющейся информации.

Между тем, для определения понятия прогнозирования спроса недостаточно представлять понятие прогнозирования и знать об объекте прогнозирования – спросе на товар (товарный запас) в розничной торговле. Это происходит из-за достаточно большого разнообразия подходов к самой методологии прогнозирования спроса.

Например, существует маркетинговый подход к прогнозированию спроса, который может быть отражен в следующем определении Котлера и Келлера: «прогнозирование — искусство предвидения поведения покупателей в определенных обстоятельствах на основе анализа результатов опросов» [40]. Следовательно, оценка будущего спроса здесь видится как величина, которую можно оценить только с помощью инструментов маркетингового анализа. Для этого производится планирование опроса, выборки, формулируются части опросника, проводится непосредственно опрос и на основании его результатов делаются выводы о рыночном спросе на данный товар. Этот способ может быть очень полезен, так как гибко выявляет предпочтения

покупателей в зависимости от изменения условий (цены, упаковки и т.п.). Тем не менее, он является достаточно затратным и не всегда оправданным для его использования при автоматизации управления товарными запасами.

Классическое определение прогнозирования спроса дает Лопатников: «Прогнозирование спроса (англ. forecasting of demand) — исследование будущего (возможного) спроса на товары и услуги в целях лучшего обоснования соответствующих производственных планов» [47]. При этом уточняется, что для осуществления прогноза используется *статистика* о реализации товаров, и, как следствие, предполагается анализ предыдущих продаж товара.

Соответственно, из определений видно, что основной сутью прогнозирования спроса является характеристика объекта прогнозирования – в данном случае спрос на товары и услуги – и подход к прогнозированию. В рамках диссертационной работы подход определяется как прогнозирование на основе статистики о предыдущих продажах и условиях, в рамках которых они осуществлялись. Здесь следует выделить пространство классических методов прогнозирования спроса: анализ временных рядов и регрессионно-эконометрический анализ и продвинутые методы машинного обучения [80, 82]. Учитывая выбранные техники, закладывается математическая природа прогнозирования спроса. При этом под математическим прогнозированием понимается формализованный вид прогнозирования, который основан на построении математической модели процесса, улавливающего большую часть его закономерностей [94].

Прежде чем приступить к процедуре прогнозирования на основании математических методов, необходимо проанализировать набор данных (X, Y) на которых будет построена модель. В рамках задачи прогнозирования спроса данные обычно состоят из набора характеристик ситуации, в которой тот или иной товар был востребован покупателем в определен количестве. Этих характеристик может быть довольно много: от погодных условий возле точки продажи до стоимости товара, на который прогнозируется спрос. В условиях большого количества данных по возможным признакам X проводится корреляционный анализ. Для начала, корреляционный анализ позволяет выявить тесноту признаков X с независимой переменной Y. Эта процедура выявляет наиболее информативные переменные для построения модели, что очень важно с точки зрения ограниченных ресурсов на использование данных. Далее, корреляционный анализ используется для выявления сильных зависимостей внутри экономических данных по признакам X, чтобы в дальнейшем использовать их с поправкой на эту зависимость или снизить их размерность путем устранения признаков, которые не несут существенной информации [44, 88].

При этом есть понимание о наличии нелинейных взаимосвязей в рамках тех или иных проявлений признаков X. Данную нелинейность можно учесть с помощью создания нелинейных комбинаций между признаками, например, процедурой попарного умножения,

полиномиального, экспоненциального, логарифмического преобразования и многих других. Обычно стремятся улучшить прогностическую способность результата, не потеряв смысл интерпретации.

Важной частью в выборе набора признаков X для прогнозирования величины Y является отсутствие логических противоречий. Например, при прогнозировании спроса, является некорректным действием использовать значение остатка на момент времени в качестве предиктора, несмотря на сильную взаимосвязь с целевой переменной. Причина тривиальна – любые колебания в конечном товарном остатке целиком и полностью зависят от реального спроса на товар, а не наоборот. В противном случае существует риск «зациклить» модель, делая ее неработоспособной. То же самое можно сказать по поводу величин, о которых нет информации на момент оценки спроса t_0. Следовательно, чтобы использовать подобные признаки, необходима разработка прогностических моделей на подобные переменные. В случае, если это выполнимо, основная модель прогнозирования может функционировать нормально или даже лучше по сравнению с более простыми реализациями [109].

1.3.1. Модели временных рядов

Наиболее часто используемым подходом в прогнозировании спроса является инструментарий анализа временных рядов. Временным рядом обычно называют последовательность измерений $y_{t_1}, y_{t_2}, \dots, y_{t_N}$ случайной величины ξ_t, которые произведены в последовательные моменты времени $t_1, t_2, \dots, t_N$ [2]. В рамках рассматриваемой задачи речь идет о временном ряде спроса на тот или иной товар. Соответственно, с помощью методов анализа временных рядов, возможно формализованное представление математической модели спроса как изменяющейся во времени случайной величины ξ_t с определенным законом распределения.

Наиболее популярными видами моделей временных рядов для прогнозирования спроса являются:

- Тренд-сезонные модели;
- Модели экспоненциального сглаживания;
- Модели авторегрессии и скользящего среднего.

Это объясняется тем, что данные модели хорошо аппроксимируют специфические элементы временного ряда спроса, например, сезонность.

Тренд-сезонные модели являются фундаментальными моделями временных рядов, которые определяют основные составляющие временного ряда: T – тренд временного ряда или так называемую тенденцию (направление), в которой движется временной ряд, S – сезонная компонента временного ряда, которая отражает амплитудные колебания вокруг линии тренда, и

E (или ε) – ошибка модели. В зависимости от характеристик процесса элементы временного ряда могут быть объединены двумя способами:

- аддитивно:

$$y_t = T_t + S_t + \varepsilon_t \quad (1.1)$$

- мультипликативно:

$$y_t = T_t \times S_t \times \varepsilon_t \quad (1.2)$$

В данном случае главной процедурой по созданию модели временного ряда является декомпозиция или разложение исходного временного ряда на составляющие T, S и E. При этом в зависимости от спецификации модели ошибка ε_t распределена следующим образом:
$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ – ошибка нормально распределена для аддитивных моделей и $\varepsilon_t \sim LogN(0, \sigma_\varepsilon^2)$ – логнормально для мультипликативных. Существует большое количество вариантов декомпозиции: классическая, декомпозиция X-12, декомпозиция с помощью создания локальных регрессий. Подробно методы декомпозиции разбираются в источнике [78]. Однако, следует обозначить, что, несмотря на довольно простую интерпретацию процесса, методы, основанные на декомпозиции, в среднем отличаются более низким качеством прогнозирования. Несмотря на это, принцип разложения ряда на составляющие используется во многих современных моделях временных рядов, что будет показано в рамках данного исследования.

Метод экспоненциального сглаживания является адаптивным методом прогнозирования. Это значит, что с получением новой информации о временном ряде, модель способна подстраиваться под изменения, тем самым повышая свою прогностическую способность. Идею экспоненциального сглаживания отражает самая простая спецификация – модель простого экспоненциального сглаживания:

$$\hat{y}_t = \alpha y_{t-1} + (1 - \alpha)\hat{y}_{t-1}, \quad (1.3)$$

где $\hat{y}_t$ – прогнозируемое значение временного ряда в момент t, y_{t-1} – фактическое значение временного ряда в момент $t-1$, α – параметр сглаживания, принимающий значения от 0 до 1. Следовательно, речь идет о процедуре «сглаживания» исходного ряда, которая позволяет увидеть изменения в тенденциях.

Существует множество модификаций исходной модели простого экспоненциального сглаживания, в том числе которые позволяют учесть компоненты временного ряда. Здесь следует остановиться наиболее подробно на подходе пространства состояний [108] (в англ. state space approach). Важной характеристикой подхода является то, что он базируется на декомпозиции временного ряда, о которой было сказано ранее. Отсюда модель экспоненциально сглаживания может быть аддитивной (1.1) и мультипликативной (1.2). Составляющие ряда (E, T, S) – будь то E, T или S – адаптируют составляющие ряда с учетом их характера. В некотором смысле

существуют два глобальных класса моделей, которые имеют собственную форму. Внутри каждого класса находятся подвиды в зависимости от адаптации конкретных компонент – каждая компонента также может быть либо аддитивной, либо мультипликативной. Аддитивный класс моделей имеет следующий общий вид:

$$\begin{cases} y_t = \omega' v_{t-l} + \varepsilon_t \\ v_t = F v_{t-l} + g\varepsilon_t \end{cases}, \quad (1.4)$$

где y_t – фактическое значение временного ряда в момент t, v_t – это вектор состояний, который содержит в себе компоненты (E, T, S), ω' - заданный измерительный вектор, F – матрица переходов, g – вектор, который содержит в себе постоянные сглаживания и ε_t – ошибка модели, распределенная нормально $N(0, \sigma_\varepsilon^2)$.

Второй класс – мультипликативные модели – описываются с помощью логарифмов, аналогично предыдущей форме:

$$\begin{cases} y_t = \exp(\omega' \log(v_{t-l}) + \log(1 + \varepsilon_t)) \\ \log(v_t) = F \log(v_{t-l}) + \log(1 + g\varepsilon_t) \end{cases}, \quad (1.5)$$

В данном выражении ошибка $1 + \varepsilon_t$ распределена логнормально $LogN(0, \sigma_\varepsilon^2)$.

Выбор лучшей модели из пространства осуществляется с помощью информационных критериев. Классически используется информационный критерий Акаике, который для класса данных моделей выражен в следующей формуле:

$$AIC = -2\log(L) + 2k, \quad (1.6)$$

где L – максимизированное значение функции правдоподобия модели, k – общее количество параметров модели. Выбор модели осуществляется по наибольшему значению информационного критерия.

Модели Бокса-Дженкинса является наиболее универсальным инструментом для прогнозирования временных рядов. Их суть заключается в объединении двух смежных процессов. Первый процесс – авторегрессии, который характеризуют зависимость текущих значений ряда от его предыдущих значений:

$$\hat{y}_t = c + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t, \quad (1.7)$$

где где $y_{t-1}, \dots, y_{t-p}$ – предыдущие фактические значения временного ряда, сдвинутые от фактического на лаг от 1 до p, c – константа модели, ε_t – ошибка модели, распределенная нормально. Общую спецификацию модели обозначают как $AR(p)$.

Моделью скользящего среднего порядка q называется модель $MA(q)$

$$\hat{y}_t = c + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}, \quad (1.8)$$

где $\varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ – предыдущие ошибки модели, сдвинутые от фактической на лаг от 1 до q [51].

Согласно методологии Бокса-Дженкинса для определения основных параметров интегрированной модели авторегрессии – скользящего среднего используется автокорреляционная (ACF) и частная автокорреляционная функции $(PACF)$. Автокорреляционная функция состоит из коэффициентов корреляции текущего значения ряда и p лагов – $corr(y_{t+p}, y_t)$. Отсюда частная автокорреляционная функция принимает следующий вид:

$$pacf(p) = \begin{cases} corr(y_{t+p}, y_t), p = 1 \\ corr(y_{t+p} - y_{t+p}^{p-1}, y_t - y_t^{p-1}), p > 1' \end{cases} \quad (1.9)$$

где y_t^{p-1} – линейная регрессия на $y_{t+1}, y_{t+2}, \dots, y_{t+p+1}$.

Исходя из характера автокорреляционных функций можно сделать выводы о параметрах моделей $ARIMA$. Сама модель имеет следующую стандартную спецификацию:

$$\hat{y}_t = c + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} + \varepsilon_t \quad (1.10)$$

По сути -моделирование заключается в корректном подборе спецификации модели $ARIMA(p, d, q)$, где d – порядок разности временного ряда, который нужен для обращения нестационарного временного ряда в стационарный. Существует два понятия стационарности временного ряда: строгая стационарность и слабая стационарность. В первом случае случайный процесс называется строго стационарным, если сдвиг во времени не меняет ни одну из функций плотности распределения:

$$f(\xi_{t_1}, \xi_{t_2}, \dots, \xi_{t_n}) = f(\xi_{t_1+\Delta}, \xi_{t_2+\Delta}, \dots, \xi_{t_n+\Delta}), \quad (1.11)$$

где Δ - целочисленное приращение к моментам времени $t_1, t_2, \dots, t_n$.

Слабо стационарным случайным процессом называется процесс, который отвечает следующим условиям: математическое ожидание постоянно и не зависит от времени $E(\xi_t) = \mu$, аналогичными свойствами обладает дисперсия – $Var(\xi_t) = \sigma^2$, автокорреляционная функция не зависит от любых других факторов, кроме как от разницы значений моментов времени [35].

Часто поиск корректной спецификации модели с помощью ACF, $PACF$ и преобразования ряда в стационарный является трудоемкой и нетривиальной задачей. Поэтому, в целях автоматизации данной рутинной процедуры существует алгоритм Хиндмана-Хандакара [117]. Обобщенно, алгоритм является итерационной процедурой, которая, изменяя параметры модели p, d и q, минимизирует значение скорректированного информационного критерия Акаике.

Информационный критерий Акаике и скорректированный информационный критерий Акаике для ARIMA характеризуются в данном случае следующими формулами:

$$AIC = -2\log(L) + 2(p + q + m + 1), \quad (1.12)$$

$$AIC_c = AIC + \frac{2(p + q + k + 1)(p + q + m + 2)}{T - p - q - k - 2}, \quad (1.13)$$

где L – максимизированное значение функции правдоподобия модели; $m = 1$, если константа по модели $c \neq 0$ и $m = 1$, если верно обратное. Данный информационный критерий разработан и используется для выбора лучшей из нескольких статистических моделей.

Соответственно, для всех ARIMA моделей в данном исследовании используется алгоритм Хиндмана-Хандакара как инструмент минимизации издержек на создание прогнозной системы. Также для отражения определенных эффектов в ARIMA моделировании используются дополнительные регрессоры. Вид изначальной функции меняется на следующий:

$$\hat{y}_t = c + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} + \omega_1 x_1 + \cdots + \omega_k x_k + \varepsilon_t, \quad (1.14)$$

где $x_1, \ldots, x_k$ – дополнительные регрессоры, введенные в модель. Здесь, могут быть использованы разные наборы признаков для улучшения качества прогноза. Это может быть, например, наличие праздника, ряды Фурье для отражения сезонности и многое другое в зависимости от контекста задачи.

Более современные модели временных рядов имеют схожие черты с рассмотренными выше. Обычно для лучшей наглядности и интерпретации результатов используют описанный подход пространства состояний, раскладывая исходный ряд на составляющие компоненты T, S и E. Каждый из элементов представляется в виде функциональной зависимости определенного вида: это может быть аддитивная линейная регрессия [122], фильтр Калмана [91] и многие другие методы, что позволяет решать разнообразный круг задач со своими уникальными ограничениями. Также возможно усовершенствование классических моделей, например, применение вместо модели простого экспоненциального сглаживания модели комплекснозначного экспоненциального сглаживания, которая показывает лучшие результаты на многих аналогичных временных рядах [120]. Очень важным развитием методов на практике является возможность использования экзогенных и фиктивных переменных внутри классического подхода на анализе самого временного ряда. Все это создает множество моделей временных рядов, использование которых приводит к высокой точности прогнозирования на распределении показателей близких к стационарным. К сожалению, в практике прогнозирования спроса на товары существует большое множество разных типов временных рядов, которые далеко не всегда удовлетворяют требованиям стационарности. Поэтому следует рассмотреть классические регрессионные модели, которые могут быть использованы для прогнозирования разнородных видов спроса.

1.3.2. Классический регрессионный анализ

Классическим методом регрессионного анализа является линейная регрессия. Собственно, метод используется не только в качестве прогнозирования, но и может позволить

отлично интерпретировать результаты модели. Отсюда, высокая значимость метода линейной регрессии в экономических и эконометрических задачах. Вид модели можно представить в матричной форме:

$$Y = b \times X + \varepsilon, \tag{1.15}$$

где Y – вектор зависимой переменной, b – вектор коэффициентов модели, X – матрица независимых переменных, ε – вектор случайных ошибок. Согласно задаче прогнозирования спроса, можно представить, что в качестве вида линейной регрессии выбирается конкретно множественная линейная регрессия. Это обусловлено большим количеством факторов, которые необходимо включить для прогнозирования в рамках столь сложного процесса. Коэффициенты модели b определяются в рамках метода наименьших квадратов, который имеет под собой основу минимизацию квадратов отклонений от фактического значения:

$$(\hat{Y} - Y)^2 \to min \tag{1.16}$$

Подробная информация о классической линейной регрессии освещается в источниках [2, 29, 42, 50, 82].

Более продвинутыми методами построения линейной регрессии являются методы с регуляризацией. Под регуляризацией имеется ввиду процедура сужения значения для некоторой группы коэффициентов b, которые оказывают меньшее влияние на качественный результат прогнозирования. Метод регуляризации обычно применяется при большом количестве коэффициентов в модели. Это актуально для задачи прогнозирования спроса, так как выборка может иметь много детерминант в зависимости от проведенного предварительного анализа. Наиболее распространенными методами, в основе которых лежит процедура регуляризации, являются гребневая и лассо регрессии [27]. При нахождении коэффициентов для гребневой или лассо регрессии минимизируют следующий функционал на основе МНК с некоторым штрафным слагаемым $\lambda \sum_{j=1}^{m} K$:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{m} K = RSS + \lambda \sum_{j=1}^{m} K \to min$$
$$\begin{cases} K = \beta_j^2, \text{если оценивается гребневая регрессия} \\ K = |\beta_j|, \text{если оценивается лассо регрессия} \end{cases}, \tag{1.17}$$

где β_j – это коэффициент модели линейной регрессии (в матричной форме – b), β_0 – свободный коэффициент (сдвиг) линейной регрессии, K – это тип оценки коэффициентов, который выбирается в зависимости от метода, λ – это гиперпараметр метода, с помощью которого выбирается сила сжатия коэффициентов модели.

Развитие эконометрического подхода к решению задач прогнозирования приводит не только к усовершенствованию регрессионных методов, но и к самому подходу в рассматриваемой задаче.

В современном мире данные имеют свойство накапливаться в огромных масштабах, соответственно появляются не только пространственные или временные выборки, но и объеденные выборки, так называемые пространственно-временные выборки или панельные данные. Обычно подобные выборки формируются для решения эконометрических проблем высокого порядка, например, данные по Российскому мониторингу экономического положения и здоровья населения НИУ ВШЭ применяется для анализа проблем в экономическом положении домохозяйств [14]. Соответственно, в сфере торговли также формируется большой объем данных. Это могут быть данные по товарам и их характеристикам за определенный период времени, что вполне отвечает требованиям пространственно-временной выборки. Для панельных данных необходим особый подход к анализу и прогнозированию на их основе, поэтому были разработаны смешанные регрессионные модели [92]:

$$y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{kji} + \sum_{k=1}^{q} d_{mj} z_{mji} + \varepsilon_{ji}, \qquad (1.18)$$

где β_0 – константа, β_k – коэффициенты для фиксированных эффектов, x_{kji} – значение фиксированных эффектов, d_{mj} – коэффициенты для случайных эффектов, z_{mji} – значение случайных эффектов. Понятия фиксированных и случайных эффектов основаны на отношении переменной к процессу исследования. Фиксированные эффекты – это то, что является управляемой переменной в системе, случайные – это переменные, влияние которых возникает случайным образом. Смешанные регрессионные модели являются по сути своей развитием стандартного метода МНК, что показывается в дальнейших разделах диссертационного исследования.

Несомненное преимущество рассмотренного подхода к задаче прогнозирования спроса состоит в том, что интерпретация полученных результатов дает представление о процессе формирования покупательского спроса в целом. Тем не менее, линейная регрессия не всегда дает качественные результаты в точности прогнозирования по сравнению с иными методами. Область научного знания – машинное обучение, которое активно уточняет и развивает методы прогнозирования, влияет в том числе и на решение задач в сфере розничной торговли. Далее описываются несколько методов прогнозирования, которые являются регрессионными по своей природе, но позволяют учесть возможную нелинейность во взаимосвязях между переменными. Они основаны на достижениях в области машинного обучения и интеллектуального анализа данных и, по сути своей, не завязываются на ту или иную область применения. Рассматриваемые

методы используются при построении исследовательского результата по прогнозированию спроса в рамках данной работы.

1.3.3. Регрессия на опорных векторах

SVM или машина опорных векторов (от англ. Support Vector Machine) является одним из алгоритмов, изучаемых в машинном обучении, который характеризуется своей универсальностью – он может быть использован как в задачах регрессии, так и в задачах классификации. Метод основан на представлении разделимости классов с помощью гиперплоскости, находящейся в $N-1$ пространстве. При этом для расширения пространства предикторов используются так называемые ядерные функции. Согласно определению, данному Воронцовым [17], функция $K: X \times X \to R$ называется ядром (от англ. kernel function), если заданная функция представима следующим образом: $K(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$ при некотором отображении $\psi: X \to H$, где H – пространство со скалярным произведением.

Расширением машины опорных векторов для задач регрессии называется метод регрессии на основе опорных векторов (от англ. Support Vector Regression). В отличие от линейной регрессии, основанной на методе наименьших квадратов, метод регрессии на основе опорных векторов использует иной функционал. Он определяется только теми остатками, которые имеют значение выше некоторой положительной константы. Далее формулируется задача построения SVR для решения задачи нелинейной регрессии [90].

Множество $\{(x_i, y_i)\}_{i=1}^{N}$ определяется как множество примеров обучения в задаче регрессии. Необходимо найти множители Лагранжа $\{\alpha_i\}_{i=1}^{N}\{\alpha_i'\}_{i=1}^{N}$, которые максимизируют целевую функцию

$$J(\alpha_i, \alpha_i') = \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i') - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i') - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j')\, K\left(\mathbf{x}_i, \mathbf{x}_j\right) \quad (1.19)$$

при наличии следующих ограничений:

$$\begin{aligned} &1. \sum_{i=1}^{N} (\alpha_i - \alpha_i') = 0 \\ &2. \begin{matrix} 0 \le \alpha_i \le C, i = 1,2, \dots, N, \\ 0 \le \alpha_i' \le C, i = 1,2, \dots, N, \end{matrix} \end{aligned} \quad (1.20)$$

где C – константа, которая задается исследователем, $K\left(\mathbf{x}_i, \mathbf{x}_j\right)$ – ядро скалярного произведения, которое, согласно теореме Мерсера, определяется следующим образом:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}\left(\mathbf{x}_j\right). \quad (1.37)$$

Два параметра в исходной постановке ε и C выбираются свободно с учетом структуры и природы данных. Указанные параметры характеризуют размерность Вапника-Червоненкиса (VC-размерность) аппроксимирующей функции:

$$F(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{N} (\alpha_i - \alpha_i') \, K(x_i, x_j), \qquad (1.21)$$

где $\mathbf{w}$ – настраиваемый вектор весов, $\mathbf{x}$ – вектор признаков.

Важным моментом является понимание типов ядерных функций, используемых при построении алгоритма. Наиболее распространенными из них являются:

- Полиномиальное (однородное) ядро: $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^d$
- Полиномиальное (неоднородное) ядро: $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^d$
- Радиальная базисная ядерная функция: $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, для $\gamma > 0$
- Радиальная базисная ядерная функция Гаусса: $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2})$
- Сигмоидальное ядро: $K(\mathbf{x}, \mathbf{x}') = \tanh(\kappa\mathbf{x} \cdot \mathbf{x}' + c)$, для почти всех $\kappa > 0$ и $c < 0$.

Выбор того или иного ядра также обусловлен типом и характером данных.

По сути машина опорных векторов по своей природе относится к универсальным сетям прямого распространения также, как и многослойный персептрон. Алгоритм получил большое распространение при решении задач экономического прогнозирования и классификации [8, 37, 77], поэтому использовать его в задаче прогнозирования спроса было бы очевидным решением.

1.3.4. Регрессия на основе метода «случайный лес»

Еще одним универсальным алгоритмом при решении задач классификации и прогнозирования является метод “случайный лес” (от англ. random forest). Алгоритм основан на построении большого количества решающих деревьев и по сути отражает простую идею о том, что усредненный результат по большому количеству простых моделей является лучшим по сравнению с результатом по одной сложной модели. Метод разработан Брейманом и Катлером [104] и по сути является ансамблем деревьев решений. Описание индуктивного алгоритма случайного леса развернуто представлено в работе Чистякова [93]:

1. Для каждого из B создаваемых деревьев решений, участвующих в ансамбле, выполняются следующие процедуры:
    - Сформировать случайную выборку с повторением S размера N по исходной обучающей выборке $D = \{(x_i, y_i)\}_{i=1}^{N}$;
    - По выборке S построить дерево решений T_i, не прибегая к процедуре пруннинга (отсечения ветвей дерева), с минимальным количеством наблюдений в

терминальных вершинах равным n_{min}. Для этого необходимо рекурсивно следовать следующему алгоритму:

i. из исходного набора n признаков случайно выбрать m признаков;

ii. из m выбрать признак, который обеспечивает наилучшее расщепление дерева. Здесь определение оптимальной расщепленности вершины дерева строится на понятии загрязненности этой вершины. Под загрязненностью понимают меру неоднородности, которая максимальна если прецеденты, связанные с этой вершиной принадлежат разным классам. Для оценки загрязненности обычно используют понятие энтропии $i(t) = -\sum_{i=1}^{c} P(\omega_j)\log_2 P(\omega_j)$,

где t – вершина дерева решений, $P(\omega_j)$ – доля примеров класса ω_j в подвыборке $D(t)$. Отсюда оптимальное расщепление определяется как $\Delta i_B(t) = \frac{\Delta i(t)}{-\sum_{i=1}^{B} P_k \log_2 P_k}$, где B – количество потомков вершины дерева t, P_k – доля примеров подвыборки $D(t)$. Здесь выбирается расщепление максимизирующее величину $\Delta i_B(t)$.

iii. расщепить выборку после нахождения оптимального разбиения на две подвыборки;

2. После проведения всех итераций на шаге 1 получено $\{T_i\}_{i=1}^{B}$ – ансамбль деревьев решений;
3. Расчет предсказания для задач регрессии по полученному ансамблю проводится следующим образом:

$$\hat{f}_{rf}^{B}(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} T_i(\mathbf{x}) \qquad (1.22)$$

Out-of-Bag – метод оценки ошибки классификации при построении случайного леса. По сути, ошибка out-of-bag оценивается на выборке неправильно классифицированных примеров без учета голосов деревьев на примерах, которые и так входят в их обучающую выборку. Метод случайного леса в виду своих особенностей является нетребовательным к форматам данных, что безусловно приводит к его достаточно широкому распространению. Одним из самых главных недостатков метода является расширенные требования к расчету и хранению количества деревьев B, которое обычно достигает нескольких сотен. Случайные леса достаточно часто применяются помимо машиного обучения как инструмент анализа данных, и в экономическом анализе: например, в страховании [24] и в анализе вероятности банкротств [23].

1.3.5. Регрессия на основе нейросетевого подхода

Ведущим алгоритмом машинного обучения, который в современном мире используется в качестве теоретической основы многих программ по прогнозированию, является искусственная нейронная сеть. Искусственная нейронная сеть есть адаптация биологического подхода к организации систем высокого порядка. Это значит, что в основе метода лежит взаимодействие нейронов – нервных клеток в живом организме. Нейроны реагируют на раздражители и передают информацию к следующим нейронам. Организация процесса в искусственной нейронной сети аналогична. Как показала практика, применение этого подхода оправдано и успешно, в особенности в задачах распознавания изображений, видеоряда и звука.

В настоящей работе рассматриваются и используются возможности многослойной нейронной сети (от англ. MLP – multi layer perceptron). По сути это однонаправленная сеть, в которой нейроны расположены на разных уровнях, соответствующих понятию слоя. При решении простых задач машинного обучения используются многослойные нейронные сети с одним скрытым слоем, как это представлено на рисунке 1.6, реже с двумя.

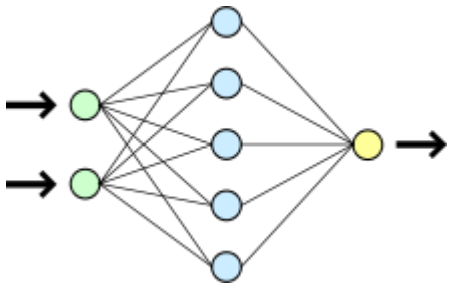


Рисунок 1.4 – Многослойный перцептрон

Существует несколько типов функционирования нейронов, т.е. базовых единиц нейронной сети. Это может быть простой персептрон, функция активации которого описывается следующим образом:

$$y_i(u_i) = \begin{cases} 1 \text{ для } u \geq 0 \\ 0 \text{ для } u < 0' \end{cases} \quad (1.23)$$

где u_i – выходной сигнал сумматорной функции

$$u_i = \sum_{j=0}^{N} w_{ij} x_j, \quad (1.24)$$

где x_j – это значение входного сигнала, w_{ij} – значение веса выходного сигнала.

Гораздо чаще при реализации нейронов в нейронной сети используют сигмоидальный вид активационной функции. Он имеет неоспоримые преимущества, так как позволяет использовать градиентные методы при обучении нейронной сети. Сигмоидальный вид функции как правило имеет 2 реализации:

$$f(x) = \frac{1}{1 + e^{-\beta x}}, \tag{1.25}$$

которую можно обозначить как униполярную функцию [61]. Биполярная функция задается с помощью тангенса:

$$f(x) = \tanh(\beta x) \tag{1.26}$$

Здесь β – параметр, который выбирается пользователем. По сути он влияет на форму кривой активации. Сигмоидальный нейрон обучается методом обучения с учителем, то есть имеется выборка примеров $\langle x, d \rangle$, где x – набор входящего вектора параметров, d – размеченный примеры напротив каждого входного набора. При этом проводится минимизация целевой функции:

$$E = \frac{1}{2}(y_i - d_i)^2, \tag{1.27}$$

где $y_i = f(u_i) = f\left(\sum_{j=0}^{N} w_{ij} x_j\right)$, x – входной вектор со значениями $x = [x_0, x_1, \dots, x_N]^T$. Здесь x_0 – это единичный вектор и $x_0 = 1$ при поляризации или $x_0 = 0$ в случае ее отсутствия.

В случае с многослойной нейронной сетью (рис. 1.6) вводятся дополнительные обозначения из-за наличия дополнительных слоев: v_j – выходные сигналы нейронов скрытого слоя. Следует определить выходные сигналы для скрытого и внешнего слоя МНС: для скрытого – это

$$v_j = f\left(\sum_{j=0}^{N} w_{ij}^{(1)} x_j\right), \tag{1.28}$$

где (1) означает номер слоя. Для внешнего слоя (выходного) соответственно:

$$y_k = f\left(\sum_{i=0}^{K} w_{ki}^{(2)} v_i\right) = f\left(\sum_{i=0}^{K} w_{ki}^{(2)} f\left(\sum_{j=0}^{N} w_{ij}^{(1)} x_j\right)\right). \tag{1.29}$$

Одним из наиболее эффективных методов обучения нейронной сети является алгоритм обратного распространения ошибки (от англ. – error backpropagation, BackProp). Он представляет собой градиентный метод обучения нейронной сети, основанной на минимизации ошибок выходов сети.

Прежде чем перейти к описанию алгоритма, следует начать с описание правила Видроу-Хоффа, которое дает понятие корректировки исходных весов сети w применяя следующее выражение:

$$\Delta w_i = \eta \delta x_i, \tag{1.30}$$

где η – это коэффициент скорости обучения, δ – ошибка сети, вычисляемая как $\delta = d - y$. Очевидно, что, зная величину изменения Δw_i, можно вычислить значение нового веса:

$$w_i(k+1) = w_i(k) + \Delta w_i, \tag{1.31}$$

где k – номер итерации обучения.

Здесь возникает проблема – ошибки на скрытом слое сети неизвестны. Соответственно возникает потребность по расчету ошибки δ на выходах скрытых слоев как общий вклад в ошибку сети. Суть метода обратного распространения как раз выражен в том, что для расчета ошибок на выходе каждого скрытого слоя и получения значений корректировки весов, необходимо передавать значение ошибки от выходного слоя к входному через все слои, как бы воспроизводя процесс обратно [60]. Опуская вывод формулы коррекции весов с учетом обратного распространения ошибки, которые описаны в [42, 49 и 50], приведем итоговое выражение:

$$\frac{\partial E}{\partial w_{ij}} = -\left(y_j - d\right) \cdot f'(S) \cdot x_i, \tag{1.32}$$

где E – целевая функция по формуле (1.43), $f'(S)$ – производная функции активации по S.

С учетом противоположного направления изменения веса от знака производной функции ошибки, выводится итоговое изменение веса Δw_{ij} с учетом значения ошибки для выходного нейрона δ_j:

$$\Delta w_{ij} = \delta_j \cdot x_i \tag{1.33}$$

Для произвольного нейрона скрытого слоя сети выходная ошибка выводится в следующем виде:

$$\delta_j = f'(S) \cdot \sum_k \delta_k \cdot w_{kj}, \tag{1.34}$$

где k – номер слоя, ошибки которого используются для расчета весов предыдущего слоя.

В рамках алгоритма и при условии сигмоидальной логистической активационной функции (1.41) вычислить ошибку нейронов как выходного, так и скрытого слоев возможно по следующим формулам:

$$\begin{gathered} \delta_j^{out} = f'(S) \cdot \left(d_j - y_j\right) = y_j(1 - y_j) \cdot \left(d_j - y_j\right) \\ \delta_j^{hid} = f'(S) \cdot \sum_k \delta_j \cdot w_{kj} \end{gathered} \tag{1.35}$$

Помимо многослойного персептрона существует большое количество иных видов нейросетевых технологий: радиальные базисные сети, рекуррентные нейронные сети, сверточные нейронные сети, нейронные сети с применение нечеткой логики. Выбор того или иного типа нейронной сети, настройка ее топологии, обычно определяется ресурсными возможностями и характеристикой самой задачи.

В целом применение нейронных сетей в задачах экономики приобрело обширный характер. В том числе нейронные сети используются и для прогнозирования уровня спроса и продаж [55]. Очень часто нейронные сети применяются для прогнозирования экономических временных рядов, что также сходно по типу с исходной задачей диссертационной работы по

прогнозированию спроса [81]. Все это делает нейронные сети методом с очень широкой областью применения. При этом они не лишены недостатков, связанных со специальной подготовкой данных и достаточно высокими требованиями на вычислительные ресурсы. Это приводит к тому, что на практике, при решении задач прогнозирования на большом объеме данных, использование нейросетевых технологий становится слишком затратным.

1.3.6. Регрессия на основе композиции (ансамбля)

Большое количество методов, которые используются при построении прогнозов, говорит о том, что нет уникального решения для любой задачи. Обычно, каждый из рассмотренных методов машинного и статистического обучения, имеет свои особенности и свое качество прогностического решения для той или иной задачи. Когда необходимо выполнить высокоточное решение задачи прогнозирования (или классификации), используется концепция композиции алгоритмов. Вводятся понятия множеств X и Y как множество признаков и множество фактических выходов (прогнозируемой и классифицируемой величины), также обозначают множество R как множество оценок. Далее рассматриваются алгоритмы, имеющие вид $a(x) = C\big(b(x)\big)$, где функция $b: X \to R$ называется алгоритмическим оператором, функция $C: R \to Y$ – решающим правилом. Отсюда выводится определения композиции алгоритмов.

Композицией T алгоритмов $a_t(x) = C\big(b_t(x)\big),\ t = 1, \dots, T$ называется суперпозиция алгоритмических операторов $b_t: X \to R$, корректирующей операции $F: R^T \to R$ и решающего правила $C: R \to Y$ [19]:

$$a(x) = C\left(F\big(b_1(x), \dots, b_T(x)\big)\right), x \in X. \qquad (1.36)$$

Основной целью закладываемой композиции является минимизация ошибок отдельных базовых моделей.

В качестве методов построения композиций или ансамблей алгоритмов машинного обучения могут считаться:

- Бэггинг (от англ. bagging). Смысл бэггинга заключается в создании большого количества случайных выборок из исходных данных простым выбором с замещением. Результатом метода считается объединение предсказаний различных алгоритмов, сделанных на сформированных случайных выборках.
- Бустинг (от англ. boosting, улучшение). Для этого алгоритма характерно последовательное улучшение результата, путем компенсирования потерь при реализации предыдущего базового алгоритма машинного обучения. Бустинг и бэггинг имеют схожую природу с рассматриваемым алгоритмом случайного леса, что говорит о генеральной идее – множество условно плохих алгоритмов при создании композиции могут дать хороший

алгоритм. Эффективность бустинга доказана на практике, в особенности бустинга, основанного на деревьях решений. Это объясняется тем, что при последовательном добавлении базовых алгоритмов увеличиваются отступы обучающих объектов.

– Стекинг (от англ. stacking). В основе этого метода создания ансамблей моделей лежит понятие мета-алгоритма над существующими базовыми алгоритмами. В простой интерпретации для реализации идеи стекинга необходимо разбить исходную выборку на две части; на первой части обучить несколько алгоритмов машинного обучения; на второй части рассчитать результат. На финальном этапе по прогнозам, полученным на последнем этапе, обучается мета-алгоритм.

Ансамбли реализуются для построения сверхточных решений, когда гораздо более важен результат прогнозирования, чем его интерпретация. В рамках задачи прогнозирования спроса этот подход имеет место быть исходя из уровня ошибки при построении базовых регрессионных алгоритмов.

1.3.7. Регрессия на основе градиентного бустинга

Градиентный бустинг является результатом развития тех идей, которые заложены в первых алгоритмах бустинга. Алгоритм начинает работу с построения начальной модели и корректирует ее, пошагово создавая последовательность деревьев регрессии или используя иные базовые методы (линейная регрессия, нейронная сеть и т.п.). Каждое дерево в последовательности создается на основании остатков модели, которая сводится на предыдущем шаге. Остатки модели по сути используются в качестве целевой переменной.

По сути решается задача:

$$F = \sum_{i=1}^{m} L(y_i, a(x_i) + b_i) \to min, \tag{1.37}$$

где $a(x_i)$ – изначально построенный алгоритм, b_i – следующий выстраиваемый алгоритм, который осуществляет корректировку ответов $a(x_i)$ до верных. Таким образом, улучшается функционал $\varepsilon_i = y_i - a(x_i)$. Рассматриваемое выражение следует определять как минимизацию функции $F(b_1, \dots, b_m)$, следовательно, необходимо выбрать эффективный метод ее минимизации. В данном случае функция от многих переменных максимально убывает в направлении своего антиградиента:

$$-\left(L'\left(y_1, a(x_1)\right), \dots, L'\left(y_m, a(x_m)\right)\right) \tag{1.38}$$

Следовательно, ответы алгоритма b_i могут быть определены следующим образом:

$$b_i = -L'\left(y_i, a(x_i)\right), i \in \{1,2, \dots, m\} \tag{1.39}$$

Настройка алгоритма происходит на обучающей выборке

$$\left(x_i, -L'\left(y_i, a(x_i)\right)\right)_{i=1}^{m} \quad (1.40)$$

Отсюда определяется и название метода – градиентный бустинг. Далее, после построения ответов алгоритма b_i, строится третий алгоритм, который корректирует указанную сумму. А затем, как уже было сказано выше, процесс продолжается. Это приводит к одному из основных параметров градиентного бустинга – количеству итераций бустинга N. Параметр влияет на точность получаемых результатов.

Следующим важным параметром является η или скорость обучения (learning rate). Суть его состоит в том, что смещение аргумента в F происходит только на часть вектора для того чтобы сохранить баланс между точностью и скоростью сходимости алгоритма:

$$a_{t+1}(x) = a_t(x) + \eta \cdot b_t \quad (1.41)$$

Параметр определен в границах $\eta \in (0,1]$.

Также важным улучшением работы градиентного бустинга является применение идеи бэггинга Бреймана. Для этого каждый новый алгоритм настраивается на подвыборке обучающей выборки размером $\lfloor \delta \cdot m \rfloor$, где δ определен на диапазоне $(0,1]$. В таком случае идет речь о применении стохастического градиентного бустинга.

Решаемая задача регрессии в рамках математического моделирования спроса решается с учетом заданной функции ошибки:

$$L(y, a) = \frac{1}{2}(y - a)^2 \quad (1.42)$$

Отсюда производная

$$L'(y, a) = -(y - a) \quad (1.43)$$

Следовательно, алгоритм $b_t(x)$ обучается на выборке

$$(x_i, y_i - a_t(x_i))_{i=1}^{m} \quad (1.44)$$

Указанная спецификация позволяет обозначить работу регрессионного алгоритма как простую поправку к ответам $a(x_i)$.

Популярной реализацией алгоритма градиентного бустинга является градиентный бустинг над решающими деревьями. Данный подход используется в рамках настоящего исследования.

1.4. Выводы

1) Описание системы товародвижения раскрывает системную сложность и экономическую проблематику на торговых предприятиях. Для улучшения работы подобных систем необходимо четко понимать структуру товародвижения и владеть инструментами экономико-математического анализа.

2) Рассматриваемые теоретические аспекты управления товародвижением и сущность покупательского спроса показывают ключевую важность качественного прогнозирования потребности для улучшения эффективности принимаемых решений. Поэтому совершенствование функционирования организации розничной торговли в рамках диссертационной работы основано на построении модели прогнозирования спроса на основе продвинутого математического и статистического анализа.
3) Отдельно выделяются и рассматриваются методы математического прогнозирования спроса для дальнейшего построения методологии прогнозирования. Исходя из цели диссертационной работы, делается вывод о необходимости применения перечисленных методов в исследуемой области знаний.

ГЛАВА 2. МЕТОДОЛОГИЯ ПРОГНОЗИРОВАНИЯ ПОКУПАТЕЛЬСКОГО СПРОСА НА ТОВАР

Современная картина применяемых методов машинного обучения, которая обозначена в Главе 1, позволяет обозначить широкий спектр применяемых технологий. Тем не менее, для решения исследовательской задачи необходимо не только знание математического аппарата, но и последовательность этапов его применения. В данной главе описывается методология прогнозирования покупательского спроса на товар с учетом наиболее эффективных методов математического прогнозирования, предназначенных для решения подобного рода задач. Кроме того, рассматривается методика предварительного экономического и эконометрического анализа данных розничных сетей, способствующих качественному результату прогнозирования и оценки его эффективности. На основании рассматриваемого анализа описывается эффективная методика конструирования переменных для математического моделирования.

2.1. Методология предварительной подготовки данных: анализ факторов влияния на покупательский спрос розничного магазина

Покупательский спрос является наиболее случайной компонентой любой модели управления запасами в торговом предприятии. Чаще всего, при недостаточном его изучении, спрос принимается как детерминированная составляющая. В общем случае это приводит к отклонениям в оценках необходимых товарных запасов. Ранее, было обозначено, что для качественного управления запасов необходима более точная оценка покупательского спроса. Поэтому для создания четкой методологии прогнозирования необходимо изучить как понятие покупательского спроса, так и круг возможных детерминант, определяющих его динамику.

На характер спроса влияет множество факторов, различающихся по природе. Например, авторы [13] выделяют следующие группы факторов:

1. Социальные: социальная структура общества, уровень развития культуры, мода, профессиональный состав населения, уровень безработицы;
2. Экономические: уровень развития экономики страны, региона, размер денежных доходов, уровень и динамика розничных цен, соотношение товарной и нетоварной форм потребления, степень обеспеченности покупателей товарами, величина налогов, культура торговли, уровень цен на коммунальные услуги;
3. Демографические: численность, половозрастной состав потребителей, размер и состав семьи, миграция населения;

4. Природно-климатические: специфика климата, национальные традиции, сезонность продаж;
5. Политические: меры по поддержке малоимущих потребителей, минимальный уровень заработной платы и прожиточного минимума, индексация доходов, международные кризисы, войны, политические конфликты;
6. Товарная политика: качество товаров, насыщенность рынка товарами, современная техника и технология производства, широта ассортимента, наличие товаров-заменителей, взаимодополняющих товаров;
7. Прочие факторы: потребительские ожидания, предпочтения потребителей, сегмент рынка, реклама.

Как видно из списка, факторов, влияющих на потребительский спрос, довольно много. Но также необходимо понимать, что на практике задача значительно усложняется:

- Затруднен поиск информации обо всех предлагаемых факторах в достаточном количестве и объеме. Это настолько сложная деятельность, что стоимость совершаемых действий при этом, может оказаться значительно выше, чем экономический эффект от создания сверхточной математической модели прогнозирования спроса;
- Внутренние бизнес-процессы розничного предприятия оказывают специфическое влияние на товарный спрос. Например, реализуемая маркетинговая программа в праздничный день может принести магазину дополнительный доход или, наоборот, снизить эффект от праздника в зависимости от качества проводимых мероприятий;
- Ограничения задачи управления запасами. Прогнозирование спроса в рамках задачи управления запасами в общем случае осуществляется на краткосрочный период. Это значит, что, во-первых, спрос оценивается на каждодневной основе, во-вторых, прогноз совершается и используется при принятии решений на горизонт не более 2-х недель или месяца. Это говорит о том, что в качестве основных детерминант могут быть использованы только те показатели, которые обновляются достаточно часто и могут являться своевременными индикаторами изменения спроса.

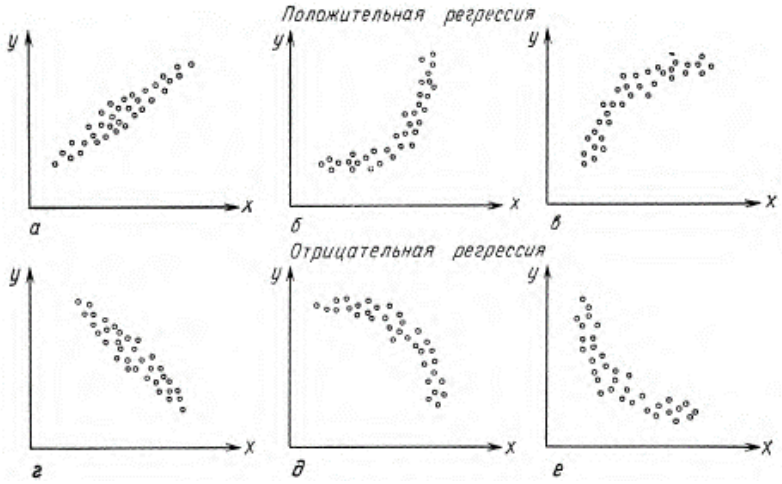
Исходя из этого, можно сделать вывод о том, что предварительный выбор факторов должен осуществляться на индивидуальной основе с учетом описанных выше особенностей. Ниже приведен список факторов, которые могут оказывать влияние на спрос и по которым имеется достаточно полная информация, а также краткое обоснование их включения в модель:

1) Внутренняя динамика продаж товара. Здесь речь идет о специфике продаж самого товара. В розничном магазине наблюдается широкий спектр товаров, которые имеют разную регулярность продаж, по-разному реагируют на дни недели и праздники, имеют особую

цикличность и сезонность. По своей сути изучение продаж определенного товара или группы товара – это изучение закономерностей самого спроса;

2) Количество покупателей в розничном магазине. Данный фактор позволяет оценить как уровень покупательской активности в магазине влияет на характер и размер спроса на тот или иной товар. Данные по чекам фиксируются в информационной системе предприятия розничной торговли и могут быть использованы для достаточно сложных видов экономико-математического анализа. В рамках корреляционного анализа будет использован только показатель общего количества за день;
3) Влияние макроэкономических параметров. Из известных макроэкономических параметров, оказывающих достаточно сильное влияние на торговлю и имеющих динамичную природу, стоит выбрать валютный курс. С учетом значительных колебаний курса, наблюдающихся с 2014 года, можно изучить влияние этого показателя на спрос и сделать выводы об его включении в модель прогнозирования;
4) Влияние календарных праздников. Наступление праздников, предположительно, имеет значительное влияние на размер товарного спроса, а также на состав покупательской корзины. Можно предположить, что накануне 23 февраля активнее ведутся продажи крепкого алкоголя или мужской парфюмерии, в то время как перед 8 марта покупатель скорее возьмет вино, конфеты и духи;
5) Влияние погодно-климатических условий. Погодный фактор сказывается больше всего на товарах повседневного спроса. Учитывая характер данных применяемых в исследовании, использовать данный фактор – обоснованная мера. В качестве иллюстрации влияния климатических факторов, можно, допустим, предположить, что в жаркую погоду растут продажи пива и мороженного.

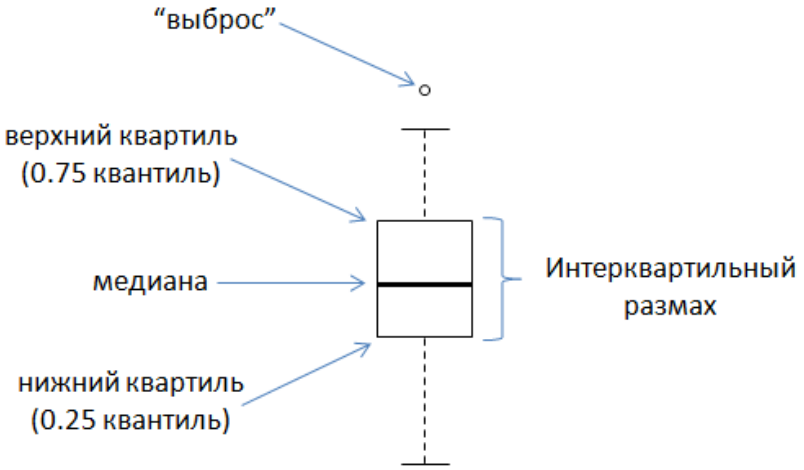
Методика выбора основных факторов, влияющих на покупательский спрос определяется с помощью корреляционного анализа [62]. В этом случае предполагается цельный анализ рассматриваемых данных, который включает в себя графический и количественный анализ зависимостей. В качестве графического анализа зависимостей выбирается диаграмма рассеяния для количественных факторов:



Источник: Фёрстер Э., Ренц Б. Методы корреляционного и регрессионного анализа: руководство для экономистов. М.: «Финансы и статистика», 1983. – 304 с.

Рисунок 2.1 - Пример диаграмм рассеяния

Для категориальных факторов – диаграмма «ящик с усами»:



Источник: Блог «R: Анализ и визуализация данных». Базовые возможности R: диаграммы размахов. URL: http://r-analytics.blogspot.ru/2011/11/r_08.html#.WnXrX4jFLIV

Рисунок 2.2 - Диаграмма «ящик с усами»

Количественной оценкой связи считается корреляционный коэффициент Пирсона, являющийся стандартным инструментом для оценки связи подобного рода:

$$r_{x_i y} = \frac{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 \sum_{j=1}^{n}(y_j - \bar{y})^2}} = \frac{cov(x_i, y)}{\sqrt{s_{x_i}^2 s_y^2}}, \quad (2.1)$$

где x_i - i-й фактор, который предположительно имеет влияние на покупательский спрос, y – целевая переменная для прогнозирования, покупательский спрос. Анализируется набор коэффициентов $r_{x_i y}$ и делаются предварительные выводы о взаимосвязи с покупательским спросом.

Подобный анализ позволяет сделать выводы о представленных данных, связанных с товародвижением в розничной торговле, и использовать эту информацию в качестве переменных в модели прогнозирования напрямую. При этом сопутствующий корреляционному анализу, содержательный анализ переменных позволяет отсеять ряд данных, которые не соответствуют постановке задачи исследования. Результатом этого этапа считается набор первоначальных гипотез о составе переменных и выводы о факторах, которые влияют на итоговый выбор товара покупателем.

2.2. Методология предварительной подготовки данных: алгоритм эвристического поиска итоговых переменных для модели прогнозирования

На основании проводимого анализа факторов и состава данных в методологии прогнозирования рассматриваемого исследования выделяется ряд переменных. При этом применяется алгоритм эвристического поиска итоговых переменных для модели прогнозирования [63]. Он основан на следующих принципах:

- Подготовка корреляционного анализа факторов влияния на покупательский спрос. Приводится список переменных x_i, который в будет оцениваться в рамках эвристического поиска;
- Разрабатывается базовая модель, с помощью которой оценена возможность включения той или иной переменной в итоговые модели прогнозирования. Для детального исследования взаимосвязей в качестве спецификации базовой модели выбрана множественная линейная регрессия общего вида:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon, \qquad (2.2)$$

где m – количество рассматриваемых переменных, y - целевая (зависимая) переменная, $x_1, x_2, \dots, x_m$ - независимые переменные и $\beta_1, \beta_2, \dots, \beta_m$ – коэффициенты при зависимых переменных, рассчитанные методом наименьших квадратов, β_0 – свободный коэффициент модели, ε – случайная ошибка модели.

Метод множественной линейной регрессии выбран исходя из простоты используемого метода в рамках задач статистики. Кроме того, он позволяет сосредоточиться на задаче прогнозирования и качественного выбора переменных для дальнейшего их использования при разработке более сложных математических моделей. Спецификация базовой модели меняется от заданной аддитивной (2.2) и сочетает в себе также нелинейные зависимости, так как это приводит к улучшению качества модели.

Изменения в рассматриваемой спецификации основаны на методе эвристического поиска минимума такого критерия качества модели как средняя квадратическая ошибка (MSE):

$$MSE = \frac{1}{n} \times \sum_{i}^{n} (y_i - \hat{y}_i)^2 \rightarrow min, \quad (2.3)$$

где y_i – фактические значения спроса, $\hat{y}_i$ – оценка прогнозной величины, n – количество элементов в выборке. MSE рассчитано строго на результатах тестовой выборки.

- Обучающая и тестовая выборки формируются из исходной в пропорциях 80% и 20% соответственно, разделяя ее на 2 временных интервала. Далее, выстраиваются первичные гипотезы о составе и типе данных, рассчитываются коэффициенты модели при данном наборе данных, затем оценивается коэффициент MSE на тестовой выборке. Проводится сравнение MSE базовой модели с аналогичным значением, которое получают с помощью простого алгоритма взвешенной скользящей средней на базе временных лагов. Использование алгоритма скользящей средней для сравнения с базовым прогнозом обусловнено тем, что он является эквивалентом метода прогноза по средней $\bar{y}$ с учетом эффекта недельной и месячной сезонностей. Недельная и месячная сезонность четко выражена для товаров повседневного спроса. Это позволяет сделать отправную точку сравнения более корректной для указанной задачи. В рамках исследования алгоритм скользящего среднего выглядит следующим образом:

$$y_t = 0.4 \times y_{t-7} + 0.3 \times y_{t-14} + 0.2 \times y_{t-21} + 0.1 \times y_{t-28}, \quad (2.4)$$

где y_{t-7}, y_{t-14}, y_{t-21} и y_{t-28} – значения товарного спроса за 7, 14, 21 и 28 дней соответственно; 0.4, 0.3, 0.2 и 0.1 – коэффициенты модели для взвешивания предыдущих значений с точки зрения усиления влияния новой информации по отношению к старой.

Дополнение и преобразование переменных производится до того момента пока

$$MSE_{lm} < MSE_{MA}, \quad (2.5)$$

где MSE_{lm} – является средней квадратической ошибкой на тестовой выборке для применяемой базовой линейной регрессии, MSE_{MA} – средняя квадратическая ошибка для алгоритма, который существует и используется на предприятии.

- Конечная цель алгоритма состоит в том, чтобы исследователь сформировал переменные, которые наиболее полно отражают анализируемые процессы. На основании этого разрабатывается базовая модель прогнозирования с категорией качества выше качества модели взвешенного скользящего среднего.

Алгоритм можно представить с помощью блок-схемы на рисунке 2.3.

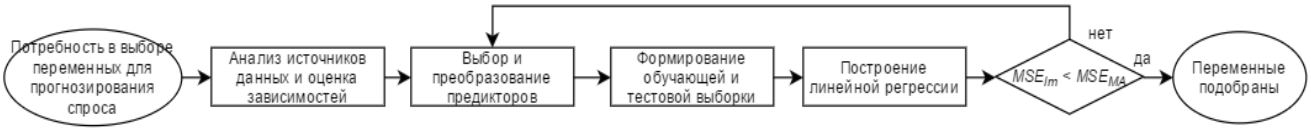


Рисунок 2.3 – Алгоритм эвристического поиска и преобразования независимых переменных

На основании работы алгоритма на рисунке 2.3 формируется ряд переменных, описанных ниже.

Код (идентификатор) и наименование товара. Единица товарной номенклатуры определяется понятием Stock Keeping Unit (SKU) или идентификатор товарной позиции. Переменная носит характер метки. У каждого SKU свой уникальный код и уникальное наименование. Корректное нахождение взаимосвязей между товарами улучшает качество прогноза.

Дата (период). Единица времени (периода) подходящего для решения сформулированной задачи – это день. Выбор в качестве периода дневных срезов обусловлено вопросами планирования товарооборота, пополнения и управления товарными запасами. На практике в розничной сети товаров повседневного спроса подавляющее большинство поставщиков розничной сети имеют график поставки товара не чаще 1 раза в день. Соответственно, нет смысла формировать более детальную развертку по часовым периодам, так как это создает дополнительные вычислительные затраты и определенные шумы в данных. Стоит отметить, что исходная выборка делится на обучающие (*training*) и тестовые (*testing*) данные по метке времени. Четкая привязка разделения выборок к времени обусловлена тем, что задача связана с прогнозом значений спроса на будущие периоды. Следовательно, на прогноз может влиять наличие нестационарности и структурных сдвигов, зафиксировать влияние которой возможно только при подобном типе разделения на обучающую и тестовую выборки.

День недели, календарные праздники и прочие характеристики временного периода. В качестве дополнительных переменных, используемых при моделировании спроса, включаются факторные переменные дня недели, наличия календарных и иных праздников (например, 14 февраля) и предпраздничных / постпраздничных дней, номер дня в праздничном периоде, номер дня в году. Данные календарные показатели позволяют зарегистрировать значимые события, влияющие на покупательский спрос.

Температурный режим. В зависимости от тех или иных погодных условий возможно и изменение спроса на товар потребительского рынка. Например, вероятен рост продаж мороженого в жаркое время по сравнению с холодным. Следовательно, переменная средней температуры воздуха в месте нахождения торговой точки (город) включена в общую структуру данных.

Количество чеков в магазине. Переменная вводится как основной фактор масштаба возможного спроса товарной группы. Очевидно, что при росте потребительского потока в общем

случае спрос на отдельные товары растет из-за разнообразия покупательских предпочтений, находящихся в магазине.

Суммарные продажи товаров в рамках одной группы. Переменная отражает масштабы продаваемого товара в рамках общей товарной группы. Это является хорошей характеристикой для будущей модели, так как общие темпы продаж группы несут в себе дополнительную информацию для расчета эффективного прогноза по спросу на конкретный товар из этой группы.

Остатки товара на начало и на конец дня. Переменные, которые выражают количество товара, находившегося в магазине на момент его открытия и закрытия. Эти данные напрямую не участвуют в моделировании целевой переменной. Тем не менее они задают дополнительные условия к ее качеству, которые будут рассмотрены ниже.

Продажи товара. Количество проданного товара является первообразной переменной по отношению к моделируемой переменной товарного спроса. Для выявления из продаж товара непосредственно значений *спроса* выполняются следующие условия:

$$y_i = \begin{cases} NA, \text{если } \varphi_i \leq 0 \\ s_i, \text{если } \varphi_i > 0 \end{cases}, \quad (2.6)$$

где y_i – величина покупательского спроса на -й товар, φ_i – минимум из значений величины остатка товара на начало дня и на конец дня, взятых из информационной базы предприятия (может быть отрицательным), s_i – значение величины продаж товара, NA – неопределенное значение целевой переменной [6].

Присвоение отсутствующих значений для целевой переменной y, говорит о том, что необходимо либо заменить отсутствующие значения с помощью какого-либо приближенного алгоритма, либо исключить данные в связке с отсутствующей переменной из исходного набора. Было принято решение об исключении данных ввиду искажения и сложности замены для целевой переменной.

Предыдущие значения спроса (лагированный спрос). Важной частью модели является включение в независимые переменные лагированные значения спроса. В продажах потребительских товаров очевидно влияние предыдущих периодов, а значит наличие автокорреляции [87]. Тем не менее выбор включаемых в модель лагов является нетривиальной задачей с точки зрения конечного качества модели.

В качестве примера можно рассмотреть график частной автокорреляционной функции одного из товаров (выбран по SKU) исходной выборки на рис. 2.4.

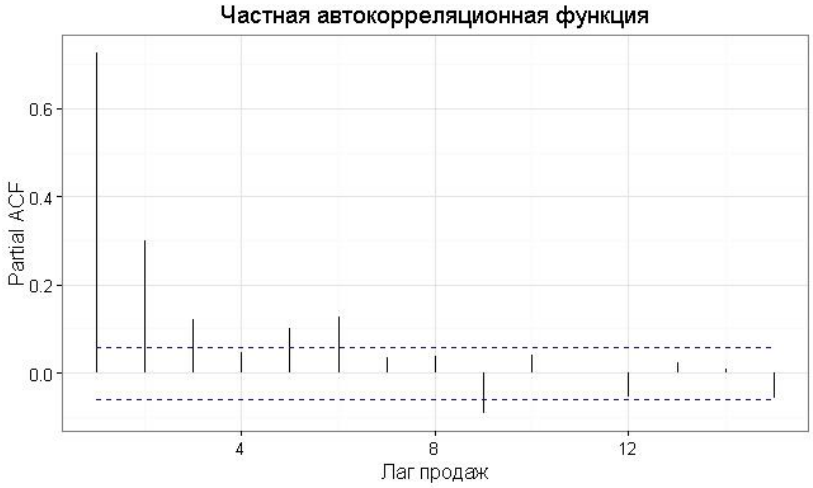


Рисунок 2.4 – Алгоритм эвристического поиска и преобразования независимых переменных

Как видно, для продаж потребительских товаров характерна затухающая структура автокорреляционных эффектов, при этом имеет смысл использовать первые семь лагов в моделировании переменной. Авторегрессионная составляющая введена на частных основаниях, поэтому вид общей линейной модели выглядит следующим образом:

$$y = \beta_0 + (\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_k y_{t-k}) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon, \quad (2.7)$$

где y - целевая (зависимая) переменная, $y_{t-1}, y_{t-2}, \dots, y_{t-k}$ – лагированные значения ряда, $\alpha_1, \alpha_2, \dots, \alpha_k$ – авторегрессионные коэффициенты модели, m – количество независимых переменных, $x_1, x_2, \dots, x_m$ - независимые переменные и $\beta_1, \beta_2, \dots, \beta_m$ – коэффициенты при зависимых переменных, рассчитанные методом наименьших квадратов, β_0 – свободный коэффициент модели, ε – случайная ошибка модели. Состав включаемых лагируемых переменных может меняться в зависимости от условия максимизации качества модели по алгоритму на рис. 2.4.

Ценовые показатели товара. Одной из ключевых характеристик товара является его цена. Стоимость товара задает меру эластичности спроса, воспринимаемое качества товара со стороны покупателя и прочие факторы, влияющие на уровень потребления. При этом, включается стоимость товара на момент времени (в выборке – день), а также – ценовой дисконт (скидка на товар) в том случае, если на момент продажи в магазине действует акция со снижением цены. Уровень скидки рассчитывается следующим образом:

$$l = \left| \frac{c_b - c_p}{c_b} \right|, \quad (2.8)$$

где l – уровень скидки, c_b- цена товара в предыдущий (базовый) безакционный период, c_p – цена товара в период промо-активности.

Наличие ценовой акции. Факторная переменная, которая выражает в себе наличие ценовой акции на товар. Это сопровождается изменениями в процедуре продажи товара: переоценка,

печать и замена обычных ценников на так называемые «желтые» ценники и вероятное расширение выкладки товара на полочном пространстве магазина. Подобное продвижение товара, очевидно, стимулирует дополнительные продажи товара.

Также в качестве отражения «эффекта памяти» у покупателя в исходную выборку добавляется лагированная переменная «Наличие ценовой акции», которая позволяет выровнять дисбаланс в оценке спроса при смене «желтого» ценника на обычный и наоборот, так как имеет место быть завышенный спрос сразу после отмены акции и заниженный при ее начале (информационный лаг).

Вес (емкость) товара. Одна из ключевых товарных характеристик, которая определяет размерность конкретной товарной позиции для покупателя. Исходя из веса (емкости) товара и иных параметрических характеристик покупатель принимает решение о количестве закупаемой продукции.

Наименование производителя и страны производства. Для определенных групп товаров ключевую роль для выбора товара потребителем играет страна-производитель и основной бренд (наименование производителя). Вводятся пространства фиктивных переменных, которые отражают принадлежность товарных позиций к производителю и к стране – $X_{M_{k-1}}$ и $X_{C_{m-1}}$ соответственно, где k – количество производителей (*makers*) в выбранной товарной группе и m – количество стран (*country*) в выбранной товарной группе. Количество переменных уменьшено на 1 с целью устранения явления «ловушки фиктивных переменных» [30, 50].

При учете данных факторов также встает задача по минимизации количества фиктивных переменных без потери информационной составляющей. В ходе этой процедуры выделяются основные производители и основные страны по исходной выборке, остальные переменные с околонулевой дисперсией – удаляются [112]. Данная процедура повышает предсказательную способность и устойчивость модели на тестовой выборке.

Как было показано, результатом работы эвристического алгоритма является целый ряд переменных: от простых значений продаж до характеристик промо-компаний на указанный товар. Данные переменные прошли четкий отбор с учетом целевой метрики алгоритма, а также экономических соображений о сущности исследуемого процесса.

2.3. Конструирование новых переменных: использование продвинутых подходов

Помимо банального отбора переменных из исследуемых данных также используются методы получения новой информации более продвинутым способом. Речь идет о процедуре создания новых переменных с использованием уже имеющейся информации об объекте без

подключения дополнительных источников данных. Чаще всего, в машинном обучении под этим подразумевается подход «feature engineering» или «проектирование признаков». В целях настоящего исследования применяется также эконометрический подход – новое измерение объекта вводится на основании экономической целесообразности измерения. Ниже представлены сконструированные на основе существующих признаков переменные.

Количество SKU, взаимозаменяемых по цене. Количество товарных позиций, взаимозаменяемых по цене – это количество складских учетных единиц, сходных по цене с товаром, по которому моделируется спрос. Для того, чтобы рассчитать количество взаимозаменяемых позиций N_i было определено k ценовых интервалов $(pr_i, pr_i + h]$, где pr_i – цена товара (нижняя граница i-ого интервала), h - шаг цены, формируемый исходя из следующего принципа:

$$h = (pr_\alpha - pr_{min}) \times d, \quad (2.9)$$

где pr_α – α–квантиль распределения цен группы товаров (в рассматриваемой задаче вероятность α принимается равной 0,9), pr_{min} – минимальное значение цены по группе товаров, d – коэффициент максимально возможного изменения взаимозаменяемых цен, выраженный в процентах. Значение pr_α использовалось при определении шага для объединения экстремально высоких значений цен в группу «pr_α и выше», т.е. для определения $k + 1$ ценового диапазона. Количество товарных позиций, взаимозаменяемых по цене N_{SKU} – это количество складских учетных единиц, принадлежащих к одному ценовому диапазону с товаром, спрос которого моделировался: если $pr_{SKU} \in (pr_i, pr_i + h]$, то $N_{SKU} = N_i$, где pr_{SKU} – значение стоимости складской учетной единицы. Исходя из формулы (1), значение показателя N_{SKU} зависит от ширины ценового диапазона и параметра d. Значение параметра d было подобрано по децильному принципу: все номенклатурные единицы разделились на ψ равнозначных групп по объему.

Показатель N_{SKU} вводился в систему переменных для отражения предположения о том, что товары с ценой в рамках одного диапазона имеют показатель средних продаж $\bar{s}_i = S_i/N_i \to 0$ при $N_i \to +\infty$, где S_i – суммарное количество продаж товаров в рамках одного ценового диапазона $(pr_i, pr_i + h]$, N_i – количество товарных позиций внутри заданного ценового диапазона, $\bar{s}_i$ – средние продажи товаров на 1 наименование внутри i ценового диапазона.

Товарные кластеры. Так как для прогнозирования спроса используется выборка с большим количеством товаров, то возникает проблема потери информации при разбиении данных на достаточно мелкие выборки. В случае этого разбиения, моделирование целевой переменной происходит на малых данных, что приводит к большой ошибке на тестовой выборке.

В ходе реализации задачи прогнозирования спроса для ежедневного потребления набора товаров выявляются следующие характеристики:

1. Товарный набор $Y = (y_1, \dots, y_i, \dots, y_N)$ не является однородным по времени, поэтому при прогнозировании были использованы временные ряды спроса каждого товара $y_i = \big(y_i(1), y_i(2), \dots, y_i(t_i)\big)$, где t_i – момент времени измерения спроса i-ого товара.
2. Информация о двух товарах из набора может быть неравнозначной по времени обозрения. Если товар со спросом y_j является новинкой, а значения y_i относятся к товару со зрелой стадией жизненного цикла, то длина временного ряда y_i будет существенно больше y_j: $t_i \gg t_j$.
3. Набор независимых величин X, присущих как набору товаров, так и каждому товару в отдельности, также может меняться во времени: $X = \big(x(1), x(2), \dots, x(t_i)\big)$.

Исходя из указанного следует, что при достаточно большом количестве N элементов в наборе, неравномерно распределенных на всем временном промежутке T решения регрессионной задачи, качество функционала для каждой случайно взятой оценки спроса y_i начинает падать. Одним из способов преодоления этого ограничения является применение кластерного анализа к набору товаров с целью выделить определенные подгруппы похожих товаров внутри набора [70]. Для проведения кластерного анализа были использованы два типа данных о товарах в наборе:

- Данные об измерениях временных рядов по величинам спроса в наборе Y, т.е. непосредственно кластеризуемые временные ряды.
- Данные о свойствах товаров в наборе. К ним относится информация о стоимости товара, его производителе, геометрические характеристики.

Выделение кластеров внутри набора производится с помощью классических методов кластерного анализа: алгоритма k-средних и EM-алгоритма (выделение гауссовых смесей распределения). Выбор данных алгоритмов обусловлен простотой исполнения и улучшением решения итоговой задачи прогнозирования.

Набор переменных для кластеризации определен таблицей 2.1.

Таблица 2.1

Набор переменных для кластеризации

Наименование переменной	Описание переменной
weekday	День недели даты, на которую осуществляется прогноз *i*-ого товара.
Y1…Y28	Значения спроса y^i для каждого товара в наборе за последние 28 дней (не включая прогнозируемую величину). Над исходными величинами была произведена процедура $\log(y+1)$

Max, Min, Median, Mean Sd	Основные структурные показатели по спросу за последний период 28 дней продаж – максимальный, минимальный, медианный и средний спрос соответственно, а также стандартное отклонение спроса.
lin	Структурный показатель, который выражает коэффициент линейного тренда по спросу за последние 28 дней. Рассчитывается с помощью метода наименьших квадратов.
numberNull	Структурный показатель, который выражает количество нулевых значений спроса за последние 28 дней
P1…P28	Наличие / отсутствие промо-акции на товар за последние 28 дней.
price	Стоимость за 1 единицу товара.
weight	Вес товара. Здесь, вместо веса, могут быть использованы любые важные геометрические характеристики товара.
country	Страна-производитель продукта. Данные включаются в виде набора фиктивных переменных.
maker	Наименование производителя продукта. Данные включаются в виде набора фиктивных переменных.

Матрица корреляций, построенная по описываемым переменным, имеет элементы, которые близки к 1, т.е. существует мультиколлинеарность между переменными. Соответственно, это является проблемой для сходимости алгоритма кластерного анализа. Поэтому для избавления эффекта коррелированности используется алгоритм преобразования исходных данных – метод главных компонент (от англ. PCA – Principal Components Analysis). Его суть состоит в том, чтобы для конечного набора указываемых в таблице 2.1 признаков $x_1, x_2, \dots, x_m \in \mathbb{R}^n$ найти такое линейное многообразие L_k среди всех линейных многообразий, определенных $k = 0,1, \dots, n-1$, такое что сумма квадратов отклонений x_i от L_k минимальна:

$$\sum_{i=1}^{m} dist^2(x_i, L_k) \to min, \qquad (2.10)$$

где $dist(x_i, L_k)$ – евклидово расстояние от точки до линейного многообразия. После преобразования исходных признаков, процедура кластеризации становится возможной.

Ключевой метрикой оценки эффективности кластерного анализа является оценка среднеквадратического ошибки MSE итогового линейного прогноза спроса:

$$Y = AC + BX + \varepsilon, \qquad (2.11)$$

где C – матрица принадлежности товара к указанному кластеру, A и B – матрицы коэффициентов регрессионного уравнения для матрицы принадлежности кластеров и для иных переменных соответственно. Среднеквадратическая ошибка MSE рассчитывается строго на тестовой выборке, которая определена правилом деления 70 на 30 (70% попадает в обучение, 30% - в тест).

Указанная в методологии кластерная структура решения позволяет свести некоторые оценки влияния независимых факторов X на Y с учетом наличия в наборе товаров разного рода

подгрупп. При этом подбор количества алгоритма кластеризации и количества классов является нетривиальной задачей. Логично в данной ситуации использовать конечную метрику по задаче прогнозирования, но также стоит обратить внимание на саму структуру кластеров.

2.4. Прогнозирование ключевых переменных, распределенных во времени

Проведенное исследование позволило выделить ряд ключевых факторов, влияющих на спрос. Поэтому их обязательное включение как показателя в математическую модель спроса является обоснованным решением и увеличивает точность получаемого результата. Однако, здесь возникает важное обстоятельство – при прогнозировании спроса на конкретный товар, лицо принимающее решение не может знать будущих показателей для ряда переменных, которые являются ключевыми и входят в итоговую модель прогнозирования в качестве независимых. Следовательно, для расчета прогноза спроса на товар необходимо рассчитать прогноз на ключевой фактор. От точности модели прогнозирования переменной зависит точность всей системы прогнозирования в целом.

В данном разделе излагается методика построения моделей прогнозирования для трех основных факторов, которые влияют на покупательский спрос: эндогенные ключевые показатели по отношению к работе розничного предприятия – количество покупателей (чеков) в магазине и количество проданных товаров в группе; экзогенный – температурный режим. Для их прогнозирования используется единый инструментарий анализа временных рядов, в том числе современные наработки в этой области. На основе проведенного сравнительного анализа методов прогнозирования для каждого из интересующих нас факторов, выявлена лучшая модель с точки зрения установленного показателя качества и сделаны выводы об прогностической способности. Разработанная методика моэет использоваться как в общем комплексе системы прогнозирования товарного спроса на предприятии розничной торговли, так и независимо – для составления необходимых планов, выработке необходимых тактических мер в управлении розничной компании.

Учитывая характер рассматриваемых переменных, можно обозначить, что при прогнозировании используются методы анализа временных рядов с четко выраженной сезонностью. При этом состав методов будет определен современными теоретическими и практическими достижениями в этой области.

2.4.1. Методика прогнозирования временных рядов

Для прогнозирования временных рядов [64] используется:

- Модель экспоненциального сглаживания, которая классически используется в прогнозировании временных рядов также, как и ARIMA;
- ARIMA-моделирование, реализованное с помощью алгоритма Хиндмана-Хандакара;
- Модель комплекснозначного экспоненциального сглаживания. Является одной из современных разработок в области анализа временных рядов и описана в трудах [120, 121];
- Вычислительный продукт Prophet, разработанный компанией Facebook. По сути данный алгоритм является аддитивной регрессионной моделью с определенным набором надбавок для вычисления сезонности и учета кратковременных важных изменений (праздников) [122];
- Байесовские структурные временные ряды. По сути является сочетанием нескольких методологически сильных предпосылок к анализу временных рядов, основанные на теории Байеса. Метод дорабатывается и используется для прогнозирования компанией Google [124].

Большой набор используемых методов мотивируется тем, что результатом исследования является не просто выбор лучшего метода, а построение их сочетания в случае целесообразности данного решения. Далее следуют рассмотреть наиболее продвинутые методы анализа временных рядов, которые используются в построении прогнозов: модель комплекснозначного экспоненциального сглаживания, Prophet и байесовские временные ряды.

Модель комплекснозначного экспоненциального сглаживания

Модель комплекснозначного экспоненциального сглаживания берет в свою основу понятие комплексного числа $x + iy$, где x и y – числа на вещественной плоскости, а i – мнимая единица ($i^2 = -1$). В теории для строгого определения принципов моделирования вводится понятие «информационного потенциала» p_t как ненаблюдаемой составляющей временного ряда, влияющей на состояние видимых значений y_t. Две вещественные переменные объединяются в одну комплексную $y_t + ip_t$. Соответственно, общий процесс описывается следующей функциональной формой:

$$y_t + ip_t = f(Q, e_t), \tag{2.12}$$

где Q – множество комплексных значений, выбранных для моделирования комплексной переменной $y_t + ip_t$.

Объединяя идею информационного потенциала и простую модель экспоненциального сглаживания, выводится следующая формула:

$$\hat{y}_{t+1} + i\hat{p}_{t+1} = (\alpha_0 + i\alpha_1)(y_t + ip_t) + (1 - \alpha_0 + i - i\alpha_1)(\hat{y}_t + i\hat{p}_t), \tag{2.13}$$

где $\hat{y}_t$ – прогнозная оценка временного ряда, $\hat{p}_t$ – оценка значения информационного потенциала, $\alpha_0 + i\alpha_1$ – комплексный параметр сглаживания.

Также как и модель экспоненциального сглаживания, комплекснозначное экспоненциальное сглаживание имеет базовую статистическую модель пространства состояний. Ее форма выражена в следующем виде:

$$\begin{cases} y_t = l_{t-1} + e_t \\ l_t = l_{t-1} - (1-\alpha_1)c_{t-1} - \alpha_1 p_t + \alpha_0 e_t, \\ c_t = l_{t-1} \end{cases} \quad (2.14)$$

где y_t – фактическое значение временного ряда, l_t – уровень временного ряда, c_t – информационная компонента при наблюдении t. В компактной форме это можно записать следующим образом:

$$\begin{cases} y_t = \omega' v_{t-1} + e_t \\ v_t = F v_{t-1} + q p_t + g e_t \end{cases} \quad (2.15)$$

Видно, что форма аналогична форме экпоненциального сглаживания (1.4).

Модель Prophet

В отличие от ARIMA моделирования, алгоритм реализованный в системе Prophet имеет под собой другую теоретическую основу. В теории модель, которая легла в основу Prophet, раскладывается на ряд компонент:

$$\hat{y}_t = g_t + s_t + h_t + e_t, \quad (2.16)$$

где g_t – трендовая компонента, s_t – сезонная компонента и h_t – компонента, которая отражает в себе информацию по праздничным периодам и другим нерегулярным событиям. То есть суть метода близка к классическим тренд-сезонным моделям.

По сути алгоритм работает подобно обобщенной линейной модели, общая спецификация которой выражена следующим образом:

$$g\big(E(Y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m) \quad (2.17)$$

Приведенная формула своим видом напоминает стандартную множественную линейную регрессию, но основное ее отличие состоит в замене стандартных линейных компонент $\beta_j x_{ij}$ на нелинейную функцию f_j от аргумента x_{ij} [27]. Это означает, что трендовая составляющая g_t является суммой функций от временных промежутков, которые настраиваются либо вручную исследователем, либо автоматически, при определении точек изменения тренда и величины изменения скорости тренда, которые происходят в этот момент - (s_j, δ_j). При этом, в рамках данной статьи мы полагаем трендовую компоненту g_t как кусочно-линейную функцию.

Для определения сезонной компоненты s_t используются следующие процедуры:

- для существующей годовой сезонности выстраиваются ряды Фурье;
- для существующей недельной сезонности используются фиктивные переменные.

Компонента h_t определяется с помощью индикаторной переменной для каждого праздника (или значимого дня). В том случае, если изменения целевой переменной $\hat{y}_t$, приуроченные к конкретному празднику, имеют место быть в некотором диапазоне дней длинной L, то данный диапазон считается указанным праздником: $Z(t|t \in [t_k; t_{k+L}]) = 1$.

Для подгонки параметров модели используется один из иттерационных методов численной оптимизации второго порядка – алгоритм Бройдена — Флетчера — Гольдфарба — Шанно или BFGS-алгоритм [45].

Байесовская структурная модель временных рядов

Байесовские структурные временные ряды – еще один метод, который можно отнести к моделям пространства состояний. Он состоит из трех основных этапов моделирования: применение фильтра Калмана, использование «spike-and-slab» метода выбора переменных и построение байесовской модели усреднения. Наиболее подробно байесовский подход к прогнозированию временных рядов описан в теоретической работе [119]. Здесь приведено несколько общих характеристика применяемых подходов, а также последовательность их применения для выхода на нужный результат.

В данном случае структурная модель временного ряда определяется следующим образом:

$$\begin{aligned} y_t &= Z_t^T \alpha_t + e_t \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t{}' \end{aligned} \qquad (2.18)$$

где y_t – значения ряда, α_t – скрытая переменная состояния. Матрицы Z_t, T_t и R_t характеризуют известные и неизвестные параметры, оцениваемые в модели. Как видно, структура похожа на аналогичные формы, приведенные для моделей экспоненциального сглаживания.

Обычно, для работы с моделями пространства состояний применяют фильтр Калмана [99]. Фильтр рекурсивно рассчитывает прогнозное распределение $f(\alpha_{t+1}|y_{1:t})$, объединяя $f(\alpha_t|y_{1:t-1})$ вместе с y_t, при этом, используя стандартный набор формул, который логически сводится к алгоритму линейной регрессии. Процедура сглаживания Калмана преобразует выходные значения фильтра для получения распределения $f(\alpha_t|y_{1:n})$, где n – количество элементов во временном ряде для каждого момента t. Так как по предпосылкам модели все составляющие имеют гауссовскую природу, то $f(\alpha_t|y_{1:t-1})$ и $f(\alpha_t|y_{1:n})$ – это многомерные нормальные распределения со средней μ_t и дисперсией D_t. Фильтр Калмана собирает информацию о временных рядах по мере итеративного движения по списку пар (μ_t, D_t). Сглаживание Калмана используется для распределения информации о более поздних наблюдениях последовательно по более ранним парам (μ_t, D_t).

Идея «*spike-and-slab*» заключается в том, чтобы снизить количество, подаваемых на вход структурной модели, признаков. Для реализации метода вводятся несколько специальных обозначений: $\gamma_k = 1$, если $\beta_k \neq 0$ и $\gamma_k = 0$, если $\beta_k = 0$, где β_k – коэффициенты при

регрессионных признаках. Обозначают также β_γ как пространство коэффициентов β где $\beta_k \neq 0$. Отсюда «*spike-and-slab*» подход выражен в оценке априорного вероятностного распределения:

$$p(\beta, \gamma, \sigma_e^2) = p_1(\beta_\gamma | \gamma, \sigma_e^2) p_2(\sigma_e^2 | \gamma) p_3(\gamma). \quad (2.19)$$

Маргинальное распределение $p_3(\gamma)$ задается с использованием распределения Бернулли:

$$\gamma \sim \prod_{k=1}^{K} \pi_k^{\gamma k} (1 - \pi_k)^{1-\gamma k} \quad (2.20)$$

Уравнение (2.20) можно упростить, о чем подробно сказано в работе [107].

Дальнейшие обозначения, определяющие строки и столбцы матрицы, – Ω_γ^{-1} для симметричной матрицы Ω^{-1}, где $\gamma_k = 1$. Тогда условные априорные распределения $f(1/\sigma_e^2 | \gamma)$ и $f(\beta_\gamma | \sigma_e, \gamma)$ могут быть выражены условной сопряженной парой:

$$\begin{gathered} \beta_\gamma | (\sigma_e, \gamma) \sim N\left(b_\gamma, \sigma_e^2 \left(\Omega_\gamma^{-1}\right)^{-1}\right), \\ \frac{1}{\sigma_e^2} | \gamma \sim G\left(\frac{v}{2}, \frac{ss}{2}\right), \end{gathered} \quad (2.21)$$

где $G(r, s)$ – гамма-распределение со средним r/s и дисперсией r/s^2.

Определяется уравнение $y_t^* = y_t - Z_t^{*T} \alpha_t$, где Z_t^* – матрица наблюдений структурной модели с $\beta^T \mathbf{x}_t$ равным 0 (здесь и далее $\mathbf{x}_t$ и $\mathbf{X}$ – признаки, определенные алгебраически и матрично). Также определяется $\mathbf{y}^* = y_{1:n}^*$, где $\mathbf{y}^*$ – это у без компоненты временного ряда.

Совместное апостериорное распределение по β и σ_e^2, условное по γ, доступно по следующим формулам:

$$\begin{gathered} \beta_\gamma | \sigma_e, \gamma, \mathbf{y}^* \sim N\left(\tilde{\beta}_\gamma, \sigma_e^2 \left(V_\gamma^{-1}\right)^{-1}\right), \\ \frac{1}{\sigma_e^2} | \gamma, \mathbf{y}^* \sim G\left(\frac{N}{2}, \frac{SS}{2}\right), \end{gathered} \quad (2.22)$$

где достаточные статистики могут быть записаны в следующем виде:

$$\begin{gathered} V_\gamma^{-1} = (\mathbf{X}^T \mathbf{X})_\gamma + \Omega_\gamma^{-1} \\ \tilde{\beta}_\gamma = \left(V_\gamma^{-1}\right)^{-1} (\mathbf{X}^T \mathbf{y}^* + \Omega_\gamma^{-1} b_\gamma) \\ N = v + n \\ SS_\gamma = ss + \mathbf{y}^{*\boldsymbol{T}} \mathbf{y}^* + b_\gamma^T \Omega_\gamma^{-1} b_\gamma - \tilde{\beta}_\gamma^T V_\gamma^{-1} \tilde{\beta}_\gamma \end{gathered} \quad (2.23)$$

По причине сопряженности, можно маргинализировать по величинам β_γ и $1/\sigma_e^2$, чтобы получить

$$\gamma | \mathbf{y}^* \sim C(\mathbf{y}^*) \frac{\left|\Omega_\gamma^{-1}\right|^{\frac{1}{2}}}{\left|V_\gamma^{-1}\right|^{\frac{1}{2}}} \frac{f(\gamma)}{SS_\gamma^{\frac{N}{2}-1}}, \quad (2.24)$$

где $C(\mathbf{y}^*)$ - нормализированная константа, которая зависит от $\mathbf{y}^*$, но не зависит от γ.

Далее совокупность всех параметров модели, за исключением β и σ_e^2, обозначается через θ. Апостериорное распределение модели, которое описано выше, оценивают с помощью цепей Маркова методом Монте-Карло [45, с. 506–509]. Сам алгоритм разбивается на следующие шаги:

1. Симуляция латентного уровня $\boldsymbol{\alpha}$ из $f(\boldsymbol{\alpha}|\mathbf{y}, \theta, \beta, \sigma_e^2)$ с использованием метода на основе фильтра Калмана, описанного в работе [107].
2. Симуляция $\theta \sim f(\theta|\mathbf{y}, \boldsymbol{\alpha}, \beta, \sigma_e^2)$.
3. Симуляция итоговых параметров β и σ_e^2 с помощью цепей Маркова со стационарным распределением $f(\beta, \sigma_e^2|\mathbf{y}, \boldsymbol{\alpha}, \theta)$.

Таким образом, прогнозирование временного ряда осуществляется на основании оцененного апостериорного распределения, что вполне соответствует парадигме байесовского вывода.

Ряды Фурье

Ключевой теоретической предпосылкой, позволяющей в данном исследовании с некоторым приближением моделировать сезонность, является применение рядов Фурье в качестве независимой переменной x практически во всех описанных выше методах. Использование рядов Фурье является частью спектрального анализа. Суть этого анализа состоит в том, чтобы представить временной ряд как сумму определенного вида частот, называемых гармониками:

$$\hat{y}_t = a_0 + \sum_{n=1}^{k} (a_n \cos nt + b_n \sin nt), \qquad (2.25)$$

где t – является характеристикой периода временного ряда в виде значений длины окружности (например, для 12-месячной сезонности значение для января $t = 0$, для декабря $t = 11\pi/6$), k определяется количество членов в сумме ряда. По сути, здесь имеет место решение регрессионной задачи для заданного вида Фурье-функции и восстановление параметров a_n и b_n с помощью метода наименьших квадратов. Исходя из этого, следует как восстановить теоретический спектра временного ряда, так и прогнозировать значения этого спектра на будущие периоды. Для решения задачи прогнозирования ключевых показателей, при использовании рядов Фурье, важно регулировать параметр k для получения качественного результата.

Для оценки качества применяемых методов, в первую очередь, исходный ряд разделяется на обучающий и тестовый: на обучающем отрезке рассчитывается модель, на тестовом – проверяется заданная метрика качества. В случае с рассматриваемыми факторами требуется краткосрочное прогнозирование на срок не более 14-21 дня. Это значит, что, если существует история временных рядов за несколько лет, то является допустимым разбить обучающую и

тестовую выборки в соотношении 0,9/0,1. При частом регулярном пересчете параметров модели подобный подход является оправданным. Принято решение выбрать за основную метрику качества корень от среднеквадратической ошибки модели ($RMSE$):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}, \quad (2.26)$$

где $\hat{y}_t$- прогноз целевой переменной, y_t – фактическое значение целевой переменной, n – количество дней в тестовом периоде, $RMSE$ – квадрат от среднеквадратической ошибки модели в тестовом периоде.

В качестве дополнительной метрики для оценки качества выбирается средняя абсолютная ошибка в процентах ($MAPE$):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100\% \quad (2.27)$$

Данная система метрик позволяет объективно оценить точность модели на тестовой выборке и сравнить используемые методы. Кроме того, каждая из предложенных метрик позволяет интуитивно понятно оценить ошибку моделей и сделать выводы о качестве.

На основе полученных результатов прогноза для каждого показателя строится таблица типа 2.2.

Таблица 2.2

Метрики качества методов прогнозирования показателей

Метрика качества	**Метод 1**	**…**	**Метод N**
$RMSE$	$rmse_1$	$rmse_i$	$rmse_N$
$MAPE, \%$	$mape_1$	$mape_i$	$mape_N$

В таблице 2.2 указывается полная информация по качеству полученных моделей. Исходя из этого, на основании перебора выделяется лучшее сочетание полученных прогнозов. Сочетание определяется среднеарифметическим от прогнозов по выбранным результатам прогнозирования:

$$F_{mean} = \frac{\sum_{i=1}^{m} F_i}{m}, \quad (2.28)$$

где F_i – временной ряд прогнозов по -му методу, m – количество выбранных методов для усреднения, в частном случае $m = N$, то есть количеству используемых методов прогнозирования. Общим правилом выбора методов для комбинации является относительно низкая корреляция результатов между прогнозами.

Итоговое представление методики обозначено в следующей схеме на рисунке 2.5.

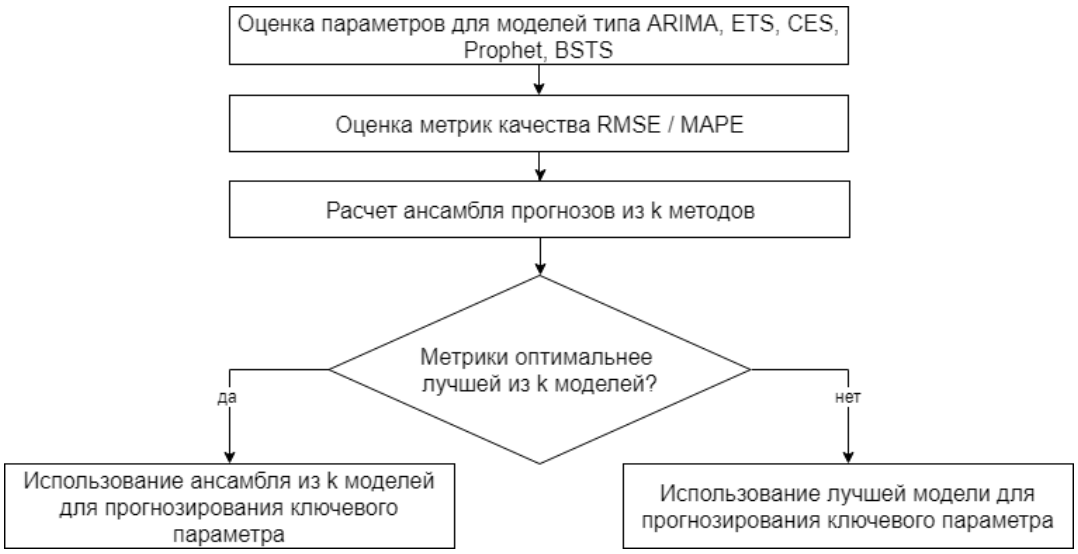


Рисунок 2.5 – Блок-схема методики прогнозирования ключевых параметров

2.4.2. Описание переменных для прогнозирования

Следует обозначить отличительные особенности временных рядов для прогнозирования:

- Количество чеков.
- Количество проданных товаров в группе.
- Температурный режим.

Количество чеков является основой оценки работы любого розничного магазина. В частности, оно характеризует интенсивность покупательского потока, который обслуживает магазин. Также показатель участвует при расчете важных финансовых метрик – например, при оценке среднего чека. В любом случае, количество чеков является обобщенной оценкой исполнения тех задач, которые ставит перед собой предприятие – покупательский поток влияет и на прибыль, и на операционные расходы, и на управление товарными запасами. По нему также делают срез успешности проведенных маркетинговых мероприятий.

Важность количества чеков для оценки около розничного (ресторанного, гостиничного) и розничного бизнеса подтверждается исследованиями в этой области: анализ показателя как определяющего объем выручки для ресторанно-гостиничного бизнеса можно найти в работах [36, 84]. Количество чеков характеризует доступность торговой точки для покупателя, а исследование [56], в котором анализируются транзакционные данные с позиции получения информации о лояльности клиентов и покупателей, расширяет понимание того, что количество чеков – это фундаментальный параметр в розничной торговле.

Для прогнозирования данного показателя применяются все рассматриваемые в исследовании методы. Кроме того, используется широкий круг экзогенных переменных:

1. Данные о наличии государственного, негосударственного и религиозного праздника за период. Для большинства методов (кроме модели Prophet, ввиду ограничений ее спецификации) используется порядковый номер дня праздника в качестве еще одного предиктора;
2. Сезонность характеризуется как годовым периодом, так и недельным. Она задается с помощь фиктивных переменных, а также рядов Фурье;
3. Учитывается также сезонность в рамках месяца, которая связана с сезонностью роста доходов населения (получение заработной платы). Подобная зависимость моделируется полиномом 5-й степени от номера дня в месяце, что также включается практически во все модели кроме Prophet.

Количество проданных товаров является важным показателем масштаба спроса на интересующие товары. С помощью него дается дополнительная информация о тенденциях внутри самой товарной группы – являются ли тренды развития восходящими или нисходящими, существует ли общая сезонность по товарной группе и т.д. Следует отметить, что прогнозируется именно количество *продаж*, а не *спрос*. Подобное допущение возможно в том случае, если товарная группа достаточно наполнена товарными позициями и каждое измерение показателя не является нулевым.

В качестве основной независимой переменной для количества проданных товаров в рамках группы используется количество чеков, так как эти два показателя связаны очевидной логикой поведения покупателей: потребитель либо планирует покупку данного товара, либо совершает его импульсно, но в отсутствии данного платежеспособного спроса каждая дополнительная покупка товара из группы невозможна. Как и для количества чеков используются:

1. Данные по праздничным дням.
2. Данные по сезонности, в том числе с помощью рядов Фурье.

Соответственно, построенный прогноз на количество проданных товаров внутри товарной группы позволяет реализовать стратегию прогноза «сверху вниз», при которой информация иерархически распространяется от более крупного агрегата к его составляющим – товарным позициям.

Влияние температурного режима на работу розничной торговли является очевидным, особенно для товаров повседневного спроса. Погодные условия влияют как на характер, так и на интенсивность покупок. Четкая сезонность спроса по многим видам товаров выделяется не столько по категориям «зима», «весна», «лето» и «осень», сколько по конкретным проявлениям погоды. Например, в жару употребляются больше освежающих напитков, в холодную погоду летом – покупательский поток ограничен для магазинов на окраине города и т.п.

Прогнозирование погодных условий, в том числе температурного режима является достаточно изученной научно-практической задачей [32]. Поэтому, при разработке системы прогнозирования товарного спроса целесообразно использовать данные из специальных сервисов по прогнозированию погоды: «Яндекс.Погода», «Google. Погода», «rp5» и других. Подобные сервисы имеют открытый API (англ. *application programming interface* – «программный интерфейс приложения») и достаточно просто интегрируются с внутренними системами предприятия. Тем не менее, существуют некоторые ограничения в их применении:

1. Достоверность прогноза погоды может быть обеспечена только на ограниченное количество дней вперед (например, для Яндекса – до 10 дней). В сочетании с тем, что цикл заказа может превышать установленные ограничения это может быть проблемой при прогнозировании спроса;
2. Использование внешних систем может привести к дестабилизации информационной системы предприятия, в том случае, если внешний сервис будет временно неисправен.

Поэтому представляется целесообразным строить внутренний прогноз температурного режима, используя описанный инструментарий анализа временных рядов. Это позволит учитывать стабильные тенденции, которые существуют при изменении температурного режима на рассматриваемой местности.

Для временного ряда температур действует несколько свойств:

- имеет место четкая годовая сезонность – это является главной компонентой при прогнозировании;
- ряд имеет сигмоидальную структуру, что позволяет успешно использовать ряды Фурье в качестве дополнительных предикторов.

Рассматриваемые ключевые переменные используются в дальнейшем в качестве составляющей в модели прогнозирования товарного спроса. Поэтому рассмотренная методика их прогнозирования в дальнейшем будет подтверждаться компьютерными вычислениями. Кроме того, в целях изучения покупательского спроса будет показана интерпретация указанных переменных в значении их влияния на покупательский спрос.

2.5. Общая методология прогнозирования спроса на товар

Важным этапом прогнозирования процессов является правильная и корректная постановка общей методологии. При этом необходимо учесть особенности моделируемого процесса и проверить предлагаемый подход экспериментально. Для этого определяется структура модели прогнозирования.

Согласно условиям задачи, модель прогнозирования спроса строится в рамках каждой товарной группы по каждому магазину розничной сети. При этом строится прогноз спроса на *каждое SKU*, так как в целях планирования и управления розничное предприятие должно контролировать продажи и запасы каждой номенклатурной позиции. Объектом исследования является розничная компания с ассортиментом товаров повседневного спроса, поэтому цикл поставки является достаточно коротким. Следовательно, прогнозирование осуществляется с шагом *один день*. Подводя итоги по структуре задачи, определяется, что система прогнозирования выстраивается на основании панельных данных. Пространственно-временная структура данных накладывает некоторые ограничения, в частности наличие цензурированных данных, а также недостаточность данных для прямого моделирования спроса на некоторые номенклатурные позиции.

Для созданной модели существует ряд допущений, которые в дальнейшем повлияли на ее научно-практическое использование:

1. Модель прогнозирования спроса является эмпирической, то есть использует историческую информацию для построения выводов об объекте моделирования.
2. Система переменных, участвующих в модели, ограничена внутренней информацией предприятия и информацией общего пользования (температурный режим, информация о календарных праздниках).
3. Модель оценивает прогноз спроса для каждой конкретной товарно-номенклатурной единицы (SKU).
4. Для построения модели используется пространственно-временная (панельная) выборка.

Иерархия моделирования представлена на рисунке 2.6 и представляет собой систему прогнозных моделей, которые выстраиваются для каждой товарной группы каждого магазина розничной сети.

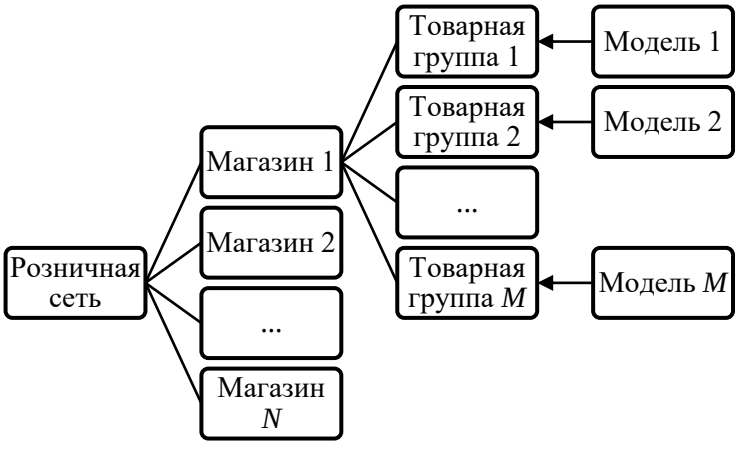


Рисунок 2.6 – Иерархия прогнозного моделирования

С учетом указанных допущений модели и иерархии выявлены основные особенности моделируемого процесса. В ходе статистических исследований обнаружено большое количество товаров с неравномерно распределенным ежедневным спросом. В теории это означает существование бимодального (мультимодального) распределения спроса на товар. Первая мода при этом находится в нуле (рис. 2.7).

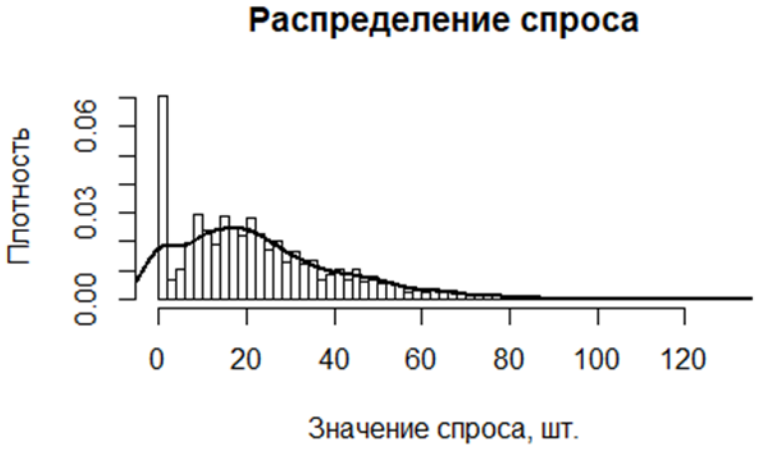


Рисунок 2.7 – Распределение спроса на товар

Подобная структура накладывает ограничение на прямое применение регрессионного подхода при прогнозировании. Последовательные эксперименты показали, что необходимо представить задачу прогнозирования спроса на товар следующими этапами [67]:

1. Определение вероятности ненулевого спроса на товар $P(y \neq 0)$. Расчет сводится к решению задачи классификации для установленного панельного набора данных;
2. Расчет прогнозного значения для всей совокупности случаев ненулевого спроса:

$$\widehat{D} = f(x|y \neq 0), \tag{2.29}$$

где $\widehat{D}$ – оценка спроса (регрессионная). Здесь речь идет о решении задачи регрессии на выборке с отсутствием значений нулевого спроса.

3. Подведение итоговой оценки прогнозного значения спроса как математического ожидания спроса на товар:

$$\hat{y} = E(y) = P(y \neq 0) \cdot \widehat{D} \tag{2.30}$$

Численные эксперименты показывают, что указанный подход является оправданным с точки зрения оптимизации целевых метрик.

Исходя из этого, решается сразу два типа задач – классификация и регрессия – а затем комбинировать полученный результат. Для решения задач применяются методы машинного обучения: логистическая и линейная регрессии с регуляризацией, случайный лес и градиентный бустинг. Нейросетевой подход и машина опорных векторов в прогнозировании исключаются из-за высоких трудозатрат при подготовке данных для данного алгоритма при достаточно низкой эффективности. Данный факт связан с тем, что в качестве входящих данных используется сильно

неоднородные показатели, имеющие разнообразную структуру. Самый важный минус, что обучение модели, при этом проходит в длительном режиме [98]. Исходя из архитектуры описанного решения на рисунке 2.5 это приводит к массовым дополнительным затратам и к потере эффективности внедряемого решения.

2.5.1. Логистическая и линейные регрессии с регуляризацией для прогнозирования спроса

Как и в случае с множественной линейной регрессией, логистическая регрессия является достаточно простым и интерпретируемым алгоритмом машинного обучения. В основе логистической регрессии лежит использование логистической функции следующего вида [38]:

$$p(X) = \frac{e^{BX}}{1 + e^{BX}}, \tag{2.31}$$

где $p(X)$ – определяемая оценка вероятности покупки товара покупателем, X – матрица независимых переменных, B – матрица коэффициентов логистической регрессии. Как видно, решение также имеет многомерную структуру, а также является линейным по сути, что можно выразить путем нескольких дополнительных преобразований:

$$\frac{p(X)}{1 - p(X)} = e^{BX} \tag{2.32}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = BX, \tag{2.33}$$

где $\log(p(X)/1 - p(X))$ – логарифм шансов. Для подгонки коэффициентов модели B используется метод максимального правдоподобия.

Состав переменных для задачи классификации в случае с логистической регрессией определяется аналогично алгоритму на рисунке 2.3. То есть оптимизируется целевая метрика, только в данном случае это AUC (англ. area under ROC curve, площадь под -кривой). Здесь и далее состав переменных для методов классификации определен в Приложении А.

Спецификации множественной линейной регрессии уже ранее была обозначена в разделе 2.2 формула (2.7).

Важно отметить, что для улучшения качества прогнозирования используется метод Elastic net (от англ. эластичная сеть), что позволяет комбинированно использовать лассо и гребневую регуляризацию. Для задачи линейной регрессии метод эластичной сети иллюстрируется следующим образом:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda\left((1-\alpha)\sum_{j=1}^{m}\beta_j^2/2 + \alpha\sum_{j=1}^{m}|\beta_j|\right) \to min, \tag{2.34}$$

где $\sum_{j=1}^{m}\beta_j^2$ – норма вектора штрафа l_2, $\sum_{j=1}^{m}|\beta_j|$ – норма вектора штрафа l_1, α – параметр смешивания двух штрафов, который принимает значения от 0 до 1. При $\alpha = 1$ имеет место быть лассо-регрессия, при $\alpha = 0$ – гребневая. Остальные параметры имеют тот же смысл, что и в (1.32). Аналогичным образом эластичная сеть строится и для логистической регрессии:

$$L(y_i|x_i,\beta_j) - \lambda\left(\frac{(1-\alpha)\sum_{j=1}^{m}\beta_j^2}{2} + \alpha\sum_{j=1}^{m}|\beta_j|\right) \to max, \qquad (2.35)$$

где $L(y_i|x_i,\beta_j)$ – оценка максимального правдоподобия.

Здесь и далее обозначается круг выбранных переменных для моделирования спроса:

- В моделировании результата по задаче определения вероятности ненулевого спроса участвуют переменные, которые можно найти в Приложении А.
- Для регрессионной задачи используются переменные из Приложения Б.

При моделировании регрессий подбираются значения коэффициентов β, а также гиперпараметры модели – α и λ. Для подбора гиперпараметров применяются техники кросс-валидации.

2.5.2. Случайный лес для прогнозирования спроса

Случайный лес как метод машинного обучения используется как при нахождении вероятности, так и при регрессионном моделировании спроса. Это показывает достаточную универсальность метода. При этом в рамках исследования используется подход вероятностного случайного леса. Его основное отличие от реализации предложенной Лео Брейманом (раздел 1.3.4) состоит в том, что выходы ансамбля $\{T_i\}_{i=1}^{B}$ формируются не по процедуре голосования большинства, а исходя из значения доли количества случаев, принадлежащих интересующему классу, в конечном узле дерева. Затем, по сути усредняются вероятности

$$\hat{p}_{rf}^{B}(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B}P_i(\mathbf{x}), \qquad (2.36)$$

где $P_i(\mathbf{x})$ – это вероятность (доля) ненулевого спроса, вычисленная для каждого -го дерева. В отличие от классического случайного леса вероятность принадлежит диапазону $P_i \in [0,1]$. Подобная формулировка актуальна для задачи классификации. Переменные, использованные для классификационного леса, отражены в Приложении А, для регрессионного – в Приложении Б.

Основной задачей при построении модели случайного леса является подбор гиперпараметров:

- N – количество деревьев в ансамбле или значение B в (2.38). При большом количестве деревьев улучается точность результата с точки зрения задаваемых метрик.

- m – количество случайно отбираемых переменных для разбиения при построении деревьев. Здесь существуют эвристики Бреймана, которые позволяют оптимально подбирать интересующий параметр [21]:

Таблица 2.3

Эвристики подбора параметра m для случайного леса

Возможные значения для классификации	Возможные значения для регрессии
$m = 0.5 \times \sqrt{M}$	$m = 0.5 \times (M/3)$
$m = \sqrt{M}$	$m = M/3$
$m = 2 \times \sqrt{M}$	$m = 2 \times (M/3)$

где M – общее количество переменных (предикторов) в данных.

Приведенные эвристики не являлись исчерпывающими при моделировании решения.

- MN – минимальное количество наблюдений в узле. Регулирует сложность формируемых деревьев, тем что косвенно определяет количество уровней дерева.

Все указанные параметры также определяются с помощью перекрестной проверки.

2.5.3. Градиентный бустинг для прогнозирования спроса

В случае применения градиентного бустинга для решения задачи прогнозирования спроса используется не только регрессионное моделирование, но и решается задача классификации. Для этого применяется логистическая функция ошибки

$$L(y, a) = \log(1 + e^{-y \cdot a}), a \in (-\infty, +\infty), y \in \{-1, +1\} \quad (2.37)$$

При разработке решения на основе градиентного бустинга использовались более продвинутые версии алгоритма, в частности, XGBoost. Его отличительной особенностью является более быстрая реализация расчетов и использование регуляризации при построении деревьев решений.

Наиболее важными параметрами при построении бустинга являются:

- B – тип базового алгоритма для бустинга: линейная модель, дерево решений, нейронная сеть и прочие. В данном исследовании применяется классический подход к определению базового алгоритма – деревья решений.
- η – скорость обучения алгоритма. Отвечает за эффективность сходимости алгоритма и его возможности попадания в глобальный минимум.
- N – число итераций алгоритма. Часто, когда в основе алгоритма принимаются деревья решений, говорят о количестве деревьев.
- MD – максимальная глубина деревьев. Отвечает за структурную сложность деревьев. Изменения параметра помогает избежать переобучения.
- MC – минимальное количество объектов в листе дерева. Также позволяет регулировать сложность деревьев.

- δ – доля объектов выборки, используемые на каждой итерации алгоритма. Позволяет увеличить устойчивость алгоритма и значительно повысить качество модели.
- c – доля используемых признаков для каждой итерации. δ и c позволяют использовать преимущества случайного леса в рамках градиентного бустинга.

Это не полный перечень возможных параметров, что говорит о достаточной степени сложности представленного алгоритма. Тем не менее, он дает качественный результат, который приводит к повышенной точности принимаемых решений.

2.5.4. Поиск гиперпараметров для методов прогнозирования спроса

Поиск всех указанных в разделах 2.5.1 – 2.5.4 гиперпараметров моделей осуществляется с помощью следующего алгоритма:

1. Определяется набор проверяемых значений гиперпараметров. Для этого используется информация о типах гиперпараметров для метода, принимаемых значениях гиперпараметров, а также о существующих эвристиках поиска гиперпараметра для того или иного метода (например, эвристики Лео Бреймана для случайного леса).
2. Для одного сочетания значений гиперпараметров проводится процедура перекрестной проверки (кросс-валидации):
    a. Исходные данные для обучения модели разбиваются на q частей. Значение q зависит от объема данных, а также вычислительной мощности.
    b. На $q - 1$ частях данных производится обучение модели при заданном сочетании гиперпараметров.
    c. На оставшейся части данных рассчитывается целевая метрика модели, которая определяет ее качество
    d. Процедура повторяется q раз, при этом каждая из частей используется для тестирования результата.
    e. Рассчитывается итоговая оценка качества модели путем усреднения всех q значений целевой метрики.
3. Процедура кросс-валидации повторяется для всех сочетаний гиперпараметров. В итоге определяется такое сочетание, которое дает наилучший результат с точки зрения целевой метрики.

Данный алгоритм также называет процедурой решетчатого поиска, так как наборы гиперпараметров образуют «решетки» сочетания значений. Алгоритм используется как в моделировании оценки вероятности ненулевого спроса, так и для регрессионного моделирования спроса.

2.5.5. Комбинация прогнозных значений спроса

Итоговым действием после моделирования результатов с помощью методов машинного обучения является комбинацией результатов. Под этим понимается, что каждый вектор прогнозов по каждому методу комбинируется в итоговый результат по каждой решаемой задаче (классификации или регрессии) как это показано на рисунке 2.8.

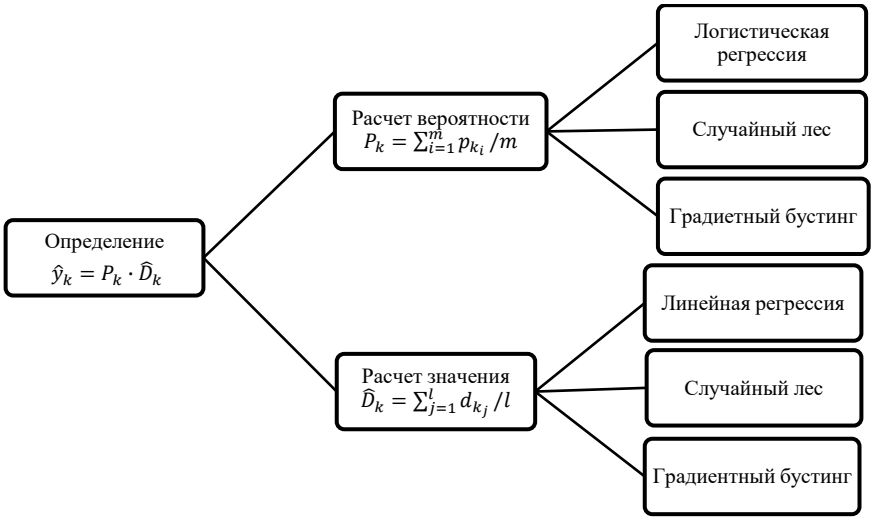


Рисунок 2.8 – Система моделирования товарного спроса

По рисунку видно, что комбинация результатов осуществляется простой процедурой среднего арифметического по всем m методам для классификации или определения вероятности ненулевого спроса и по всем l регрессионным методам. По формуле (2.30) итоговое решение выглядит следующим образом:

$$\hat{y} = E(y_k) = P_k(y \neq 0) \cdot \widehat{D}_k = \sum_{i=1}^{m} p_{k_i} / m \cdot \sum_{j=1}^{l} d_{k_j} / l \qquad (2.38)$$

где p_{k_i}– оценка вероятности ненулевого спроса, рассчитанные i-м методом, d_{k_j}– оценка регрессионного значения спроса, рассчитанного j-м методом.

Результатом применения данной методологии является теоретический и программный комплекс для оценки будущего спроса на товары в рамках одной товарной группы розничного магазина. Данный результат может быть масштабирован как на все товарные группы, так и на все магазины розничной сети, при этом этапы моделирования циклически повторяются по всем веткам иерархии, исходя из рисунка 2.5.

2.6. Метрики качества прогнозирования

Очень важным обстоятельством является определение метрик качества для итогового прогноза спроса на товар. С учетом структуры решаемой задачи, используются метрики как для задачи классификации, так и для задачи регрессии.

Согласно постановки задачи, последовательность определения качества итогового прогнозирования спроса можно обозначить следующим образом:

1. Решается задача классификации по определению вероятности ненулевого спроса. Для качества решения используется метрика AUC (от англ. Area Under Curve – площадь под кривой) кривой ROC. ROC-кривые являются универсальным инструментом оценки классификационных алгоритмов. Площадь под R-кривой при этом, достаточно просто показывает качество модели: чем она выше, тем выше дискриминирующая способность алгоритма. Указанная метрика используется для каждого метода классификации применяемого для моделирования оценки вероятности. Кроме того, при комбинации алгоритмов также происходит расчет итоговых значений метрики AUC.
2. Решается регрессионная задача прогнозирования спроса при условии, что в исторических данных нет нулевых продаж. Здесь используется ряд стандартных регрессионных метрик, которые являются адекватными при заданном условии отсутствия нулевых продаж.

$$MSE = \frac{\sum_{i=1}^{n}\left(D - \widehat{D}\right)^2}{n} \quad (2.39)$$

$$MAE = \frac{\sum_{i=1}^{n}\left|D - \widehat{D}\right|}{n} \quad (2.40)$$

   Указанные метрики также используются для оценки всех регрессионных алгоритмов. Аналогично рассчитывается оценка для комбинации алгоритмов и определяется наиболее удачная.
3. Находится итоговое решение по формуле (2.40). Исходя из состава решения, итоговыми метриками определяются MAE и MSE.

Среди всех комбинаций алгоритмов, используемых для решения задачи, выбирается та комбинация, которая наилучшим образом оптимизирует метрику на итоговой итерации. В случае наличия нескольких комбинаций с лучшим качеством выбирается та из них, которая является наиболее простой (согласно принципу бритвы Оккама).

Цикличность алгоритма определения качества представлена на рисунке 2.9.

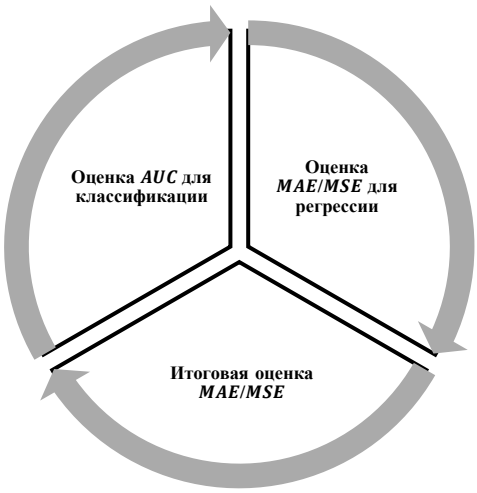


Рисунок 2.9 – Цикличность определения качества модели прогнозирования спроса

На основании итоговой метрики делаются общие выводы о качестве модели прогнозирования товарного спроса. Также принимается решение о внедрении модели в процесс принятия решений в виде реализации программного комплекса.

2.7. Выводы

1) Прописана процедура выбора факторов, влияющих на товарный спрос, затем определены переменные, которые будут участвовать в моделировании. Система переменных является базисом для дальнейшего эконометрического исследования.
2) Разработан алгоритм создания новых переменных (эконометрический анализ, кластеризация), которые позволяют эффективно улучшить решение рассматриваемой задачи: количество товарных позиций, взаимозаменяемых по цене, и товарные кластеры.
3) Разработана методика прогнозирования ключевых переменных, которые включены в реализацию итоговой модели прогнозирования. Методика основана на применение современных инструментов временных рядов. Прогноз, полученный в качестве результата применения методики, позволит избежать использования значений переменных из будущего.
4) Определена общая методология прогнозирования покупательского спроса на товар. В основе рассматриваемой методологии лежат практические выводы о природе моделируемой переменной – скученность распределения в нуле. Обозначены методы и технология машинного обучения, с помощью которых будет моделироваться результат. Раскрыты метрики качества для всей иерархии создаваемых моделей для корректной оценки результата.

ГЛАВА 3. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ МЕТОДОЛОГИИ ПРОГНОЗИРОВАНИЯ ПОКУПАТЕЛЬСКОГО СПРОСА

Построенная методология прогнозирования спроса является теоретико-практическим комплексом, который решает установленную задачу. Это накладывает определенное обязательство на практическую апробацию результатов на реальных данных торгового предприятия. Исходя из этого, было выбрано одно из предприятия розничного сектора города Ижевска, на основании задачи которой была разработана как сама теоретическая часть методологии, так и проведены необходимые вычислительные эксперименты. Данные эксперименты позволили рассчитать все указанные модели, получить результат, а также оценить его согласно задаваемым метрикам качества. Кроме того, была произведена предварительная оценка факторов влияния на спрос, произведен отбор переменных и получена исчерпывающая информация для построения моделей. Все это выразилось в готовом программном комплексе, созданном на языке программирования R, который является флагманским в решении задач математического моделирования и статистики в странах Европы и США.

В данной главе рассматривается последовательность практического применения методологии, включая информацию о конкретном программном продукте, созданном в ходе исследования. Итогом главы является экономическая оценка работы прогнозной модели с точки зрения приложения результатов прогноза в управлении заказами и товарными запасами.

3.1. Предварительная подготовка переменных для прогнозирования спроса

В качестве источника данных для проведения экспериментов выбрана торгово-розничная сеть «Гастроном», которая действует на территории города Ижевска. Основным форматом сети «Гастроном» являются торговые магазины типа супермаркет с ассортиментом товаров повседневного спроса. Временные периоды пула данных – для оценки основных факторов влияния и отбора переменных использовались данные с 01.04.2012 по 18.10.2015; для построения модели прогнозирования – 01.01.2013 по 30.09.2016. Так как, согласно методологии прогнозирования, модели должны выстраиваться в рамках цепочки «магазин – товарная группа» (см. рис. 2.5), поэтому был выбран один из магазинов розничной сети с наиболее активной историей продаж, в котором выделен ряд товарных групп для реализации методологии. Подобная структура решения позволила учесть характеристики магазина, которые заложены в информации о спросе. Например, в данных о продажах скрыто местоположение магазина – близость к основному городскому траффику, социально-демографическое положение населения района, в котором находится магазин и т.п. Все это повлияло на основные показатели потребительской

корзины, свойственной этому магазину. Набор самих данных был определен исходя из рассматриваемых гипотез о факторах влияния на спрос.

3.1.1. Корреляционный анализ факторов

Основа для любых выводов о переменных, которые необходимо включить в итоговую математическую модель – это подробный предварительный анализ исходных данных. В настоящем разделе представлен корреляционный анализ факторов в самом общем смысле – это и графический анализ зависимостей, и количественный, а также содержательные гипотезы и выводы о наличии взаимосвязи тех или иных показателей с целевой переменной.

В качестве базовых данных для оценки факторов влияния на покупательский спрос были выбраны продажи по товарным группам пиво, вино и конфеты. Выбор нескольких групп позволил сравнить результаты и подтвердить выводы о влиянии того или иного фактора на общий покупательский спрос. В тоже время, анализ разных товарных групп дал возможность увидеть отличия в динамике спроса и сделать выводы о том, что оказывает влияет на них.

Внутренняя динамика продаж товара

На графиках ниже изображена дневная динамика спроса трех выбранных для анализа товарных групп в период с 01.04.2012 по 18.10.2015 в одном из магазинов розничной сети ООО «Гастроном» в городе Ижевске, а также приведен краткий анализ динамики.

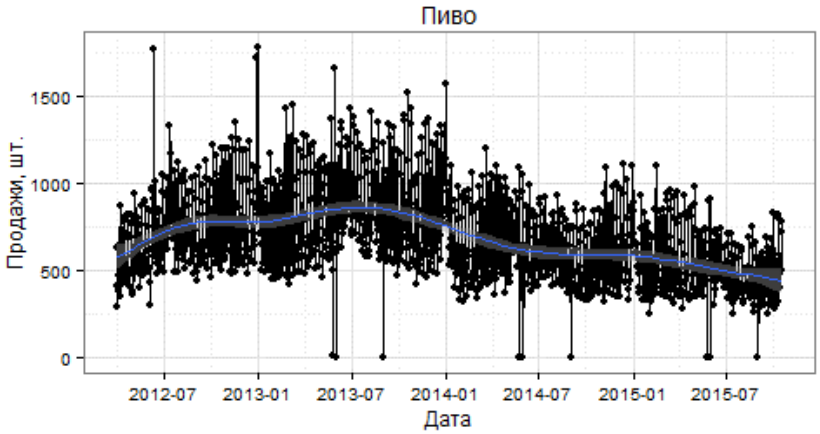


Рисунок 3.1 – Продажи группы «Пиво»

Из рассматриваемого графика видно, что дневные продажи на первый взгляд имели хаотичную структуру. Продажи менялись от минимальных 0 значений за день до 1777 бутылок – сказывались определенные внешние факторы, например, празднование нового года и запрет продажи алкоголя в определенные праздники. В целом, ясно, что тренд шел в сторону уменьшения потребления пива с конца 2013 года. Это связано как со снижением реальных доходов населения в период экономического кризиса в России, так и с законодательными инициативами.

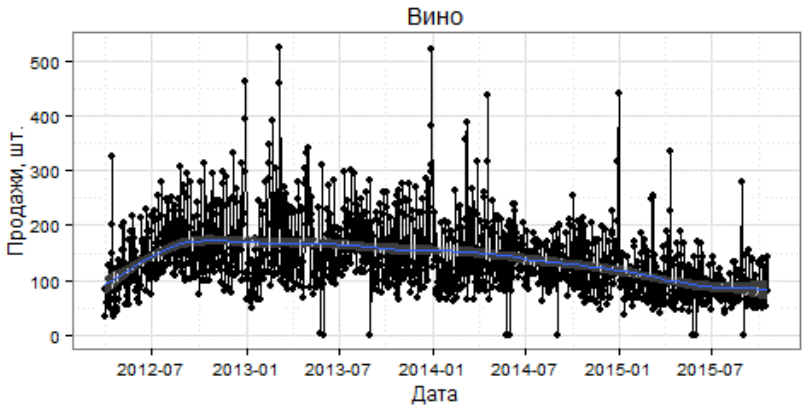


Рисунок 3.2 – Продажи группы «Вино»

По рис. 3.2 видно, что продажи группы «Вино» похожи по потребительской динамики на группу «Пиво»: тренд долгое время имел нисходящий характер, минимальные продажи за период также достигали нуля в те же промежутки времени, что и по группе «Пиво». Из отличий – продажи группы «Вино» имели меньший средний размах по сравнению с группой «Пиво» и меньший объем в количестве бутылок, что связано с иной культурой потребления данного алкогольного напитка и иными потребительскими свойствами [79].

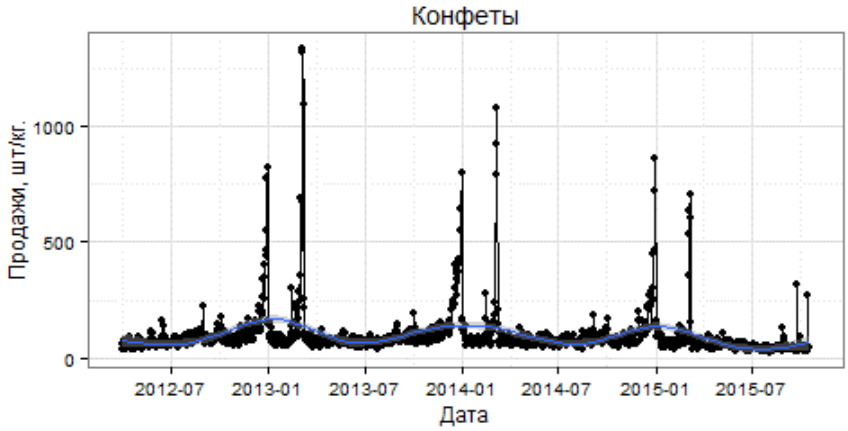


Рисунок 3.3 – Продажи группы «Конфеты»

Выводы по рис. 3.3: группа «Конфеты» имела иную специфику потребления в отличие от алкогольных групп на графиках выше. При анализе графика видно, что среднедневной размах потребления остался на более низком уровне, чем алкогольные группы (примерно до 50 штуко-килограмм в день), но при этом наблюдалась четко выраженное влияние праздников: Нового года и Международного женского дня. Рост продаж в период праздников превышал средние в несколько раз, возрастая при этом постепенно в течение от двух недель до месяца до наступления события. Минимальные продажи за период равны 24 штуко-килограммам, максимальные – 1337.

Был проведен корреляционный анализ между трех рассматриваемых товарных групп и общих продаж в магазине за аналогичный период. Данные приведены в таблице 3.1.

Таблица 3.1

Матрица корреляций между товарными группами и общими продажами

Товарные группы	Конфеты	Вино	Пиво	Общие продажи
Конфеты	1			
Вино	0.4640796	1		
Пиво	0.2641492	0.7835376	1	
Общие продажи	0.5244534	0.7966729	0.7701100	1

По таблице можно судить, что существовала достаточно тесная взаимосвязь между общими продажами магазина и остальными товарными группами. Наиболее тесная связь общих продаж наблюдалась с продажами групп «Вино» и «Пиво», менее – с продажами группы «Конфеты», всего 0,524.

Исходя из начального графического анализа были сделаны следующие выводы:

- Характер покупательского спроса во времени разнится от одной товарной группы к другой. Для детального понимания сезонности динамики товарных групп необходимы дополнительные тесты, используемые в анализе временных рядов.
- Существует некоторая цикличность в динамике продаж рассматриваемых групп. Как можно видеть данная цикличность во всех случаях равна 1 году.
- Характер тренда продаж товарных групп во многом зависит от внешних эффектов и факторов.
- Корреляционный анализ показал тесноту связи между рассматриваемыми товарными группами и общими продажами.

Для того, чтобы определения сезонности в данных о продажах товарных групп, оценивались коррелограммы и делались соответствующие выводы. Были детально проанализированы коррелограммы автокорреляционной и частной автокорреляционной функции всех трех товарных групп (рис. 3.4).

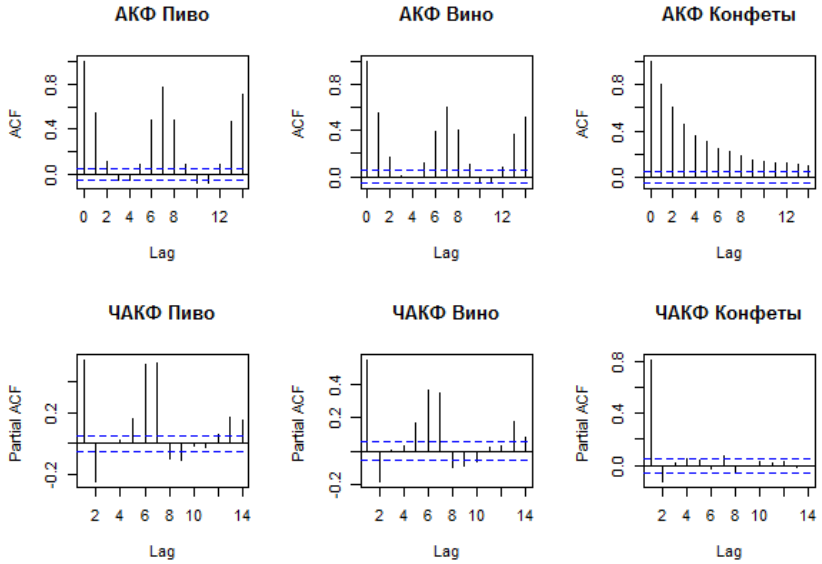


Рисунок 3.4 – Коррелограммы продаж товарных групп

Исходя из анализа коррелограмм по продажам группам «Пиво» и «Вино», были сделаны выводы о похожем характере сезонности данных временных рядов. Для лагов равных 1, 6 и 7 наблюдались максимальные значения корреляции даже после устранения зависимости между промежуточными наблюдениями. Это говорит о наличии:

- Сезонности внутри календарной недели – например, продажи в субботу с большей вероятностью отличаются по отношению к продажам в пятницу и т.д.
- Недельной сезонности – это значит, что в стандартном случае потребительская активность в текущий понедельник будет схожа с предыдущим понедельником и т.д.

Характер графика на рисунке 3.4 для группы «Конфеты» отличался от представленных алкогольных групп. У данной товарной группы не прослеживалась выраженная недельная сезонность, наиболее весомая корреляция наблюдалась при лаге 1.

Для дальнейшего моделирования использовалось предположение о наличии недельной сезонности во временных рядах продаж в товарных группах, несмотря на наличие некоторых отклонений, например, как это видно в группе «Конфеты». Это предположение связано с определенной сезонностью посещений покупателей в розничные сети и магазины, что кажется тривиальным – есть отличие между будними (рабочими) дня и выходными. Подробнее характер покупательского потока рассматривался при анализе динамики количества чеков.

Для наглядного рассмотрения предположения о влиянии определенного дня недели на продажи группы товаров рассматривался график средних продаж в раскладке по дням недели (рис. 3.5 – 3.7).

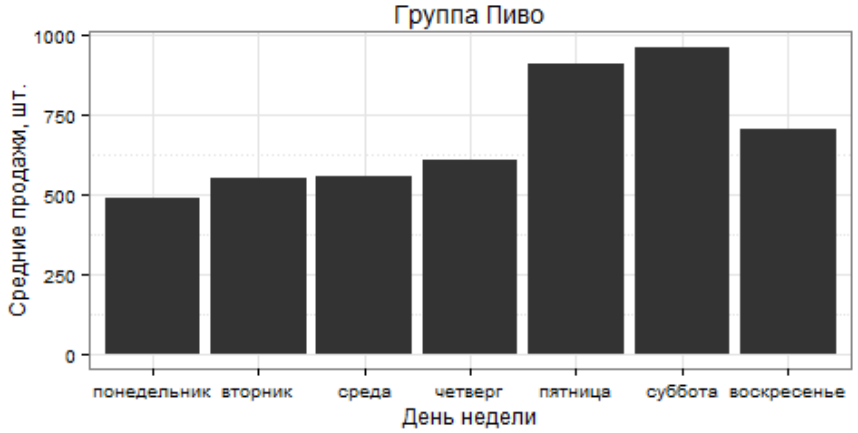


Рисунок 3.5 – Средние продажи по дням недели для группы «Пиво»

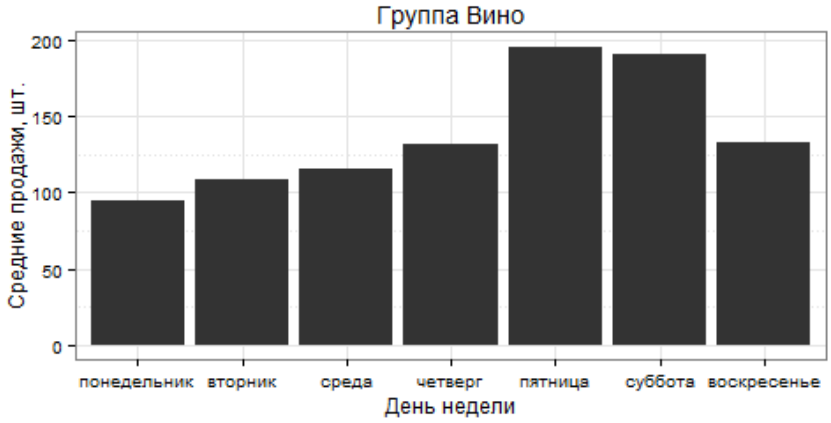


Рисунок 3.6 – Средние продажи по дням недели для группы «Вино»

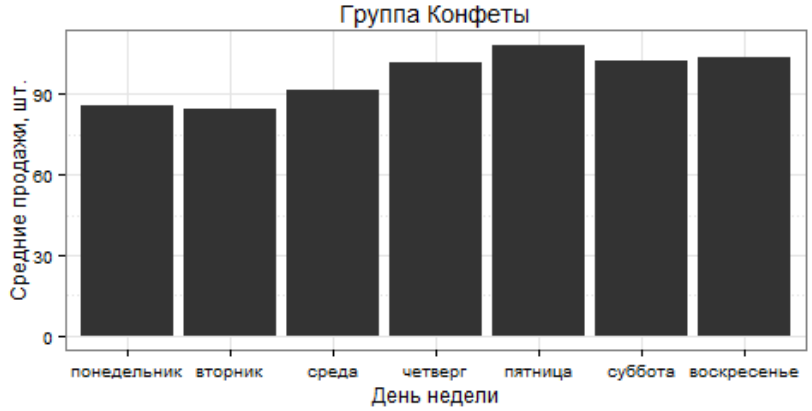


Рисунок 3.7 – Средние продажи по дням недели для группы «Конфеты»

Из графиков (рис. 3.5 – рис. 3.7) стало ясно, что сезонность внутри недели присутствовала во всех группах, в т.ч. повышение уровня продаж к наступлению выходных дней показывала и группа «Конфеты». Это подтвердило вывод о том, что необходимо использовать недельную и годовую сезонность товара как внутренний фактор при построении прогноза.

Анализ влияния количества покупателей (чеков) на продажи

Перейдем к важному аспекту рассмотрения указанных факторов – влияние количества покупателей (чеков) на продажи товарных групп в определенный промежуток времени. Взаимосвязь тривиальна, тем не менее ее количественная оценка была необходима для дальнейших этапов построения модели.

Для начала был построен график динамики чеков за период с 01.04.2012 по 17.10.2015 гг.

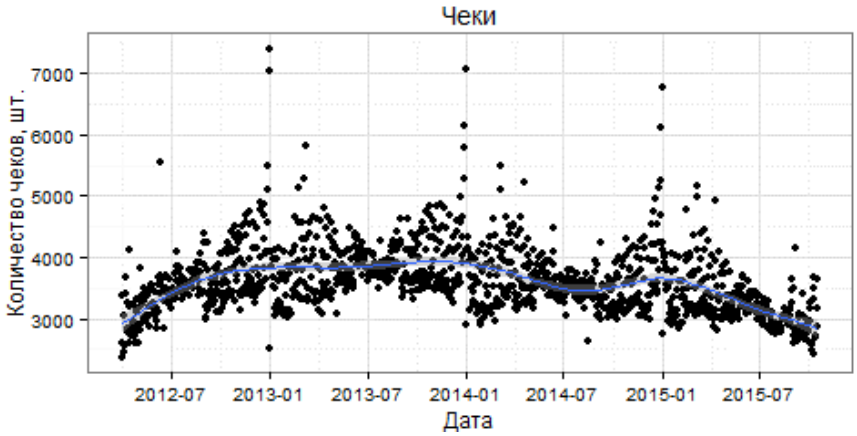


Рисунок 3.8 – Динамика количества чеков

По графику выяснилось, что динамика количества чеков также имеет схожую структуру с динамикой продаж товарных групп (рис. 3.5 – 3.7). Также имелся повышенный покупательский поток в праздничные дни, в том числе максимальные продажи в период Нового года (7402 чека 31.12.2012). Сохранялась годовая сезонность, и был виден общий тренд снижения количества покупателей.

Далее рассматривалось среднее количество чеков за анализируемый период по дням недели:

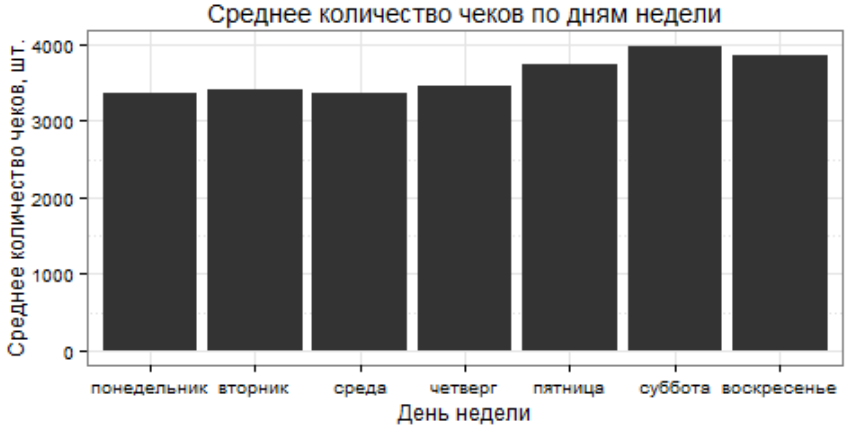


Рисунок 3.9 – Среднее количество чеков по дням недели.

По графику видно, что в динамике количества чеков существует сезонность – количество покупателей, сделавших покупку, зависело от дня недели. Это формально подтвердило сделанный выше вывод о характере сезонности посещений продуктовых магазинов.

Таблица 3.2

Матрица корреляций между продажами товарных групп и количеством чеков

Показатели	**Конфеты**	**Вино**	**Пиво**	**Кол-во чеков**
Конфеты	1			
Вино	0.4639406	1		
Пиво	0.2639680	0.7834425	1	
Кол. чеков	0.5460133	0.7326714	0.7369264	1

При оценке корреляции (Таблица 3.2) была выявлена тесная взаимосвязь между продажами групп «Пиво» и «Вино» и количеством чеков – коэффициенты корреляции 0,737 и 0,733 соответственно. Конфеты менее связаны с динамикой чеков, коэффициент корреляции равен 0,546. Можно предположить, что имеет место внесезонный характер группы. Тем не менее, видно, что показатель количества чеков важно использовать в дальнейшем моделировании спроса.

При построении модели показатель количества чеков будет браться за основу прогноза количества продаж по каждому SKU. Это связано с выводами, сделанными выше:

- Динамика и сезонность покупательского потока во многом определяет характер продаж товаров вне зависимости от товарной группы;
- При наступлении каких-то знаковых событий (праздников) покупатель чаще всего принимает решение о посещении конкретного магазина либо сначала, либо одновременно с определением необходимого списка товаров (исключением является мощная маркетинговая активность). Поэтому в данной исследовательской работе, прогноз количества чеков определяется на первых этапах моделирования спроса.

Влияние макроэкономических параметров

На работу торговой сети в значительной мере влияет общее экономическое положение как в стране, так и за рубежом. Можно выделить несколько важных факторов, влияющих на товарный спрос:

1. Реальные располагаемые доходы населения.
2. Монетарные индикаторы: ключевая процентная ставка, ставка по потребительским кредитам.
3. Индекс потребительских и производственных цен.
4. Курс иностранной валюты на отечественную и многие другие.

К сожалению, в целях создания динамичной системы прогнозирования спроса, которая работает на ежедневной основе очень сложно подобрать макропоказатели, которые сигнализируют об изменении потребления достаточно быстро. В открытых источниках информация о большинстве показателей формируется с лагом от 3-х и более месяцев, они также агрегированы по месяцам, кварталам и более широким периодам. Исключением в данном случае является такой индикатор как валютный курс.

При этом влияние валютного курса на потребительский спрос является неоспоримым – повышение международной стоимости национальной денежной единицы (падение валютного курса) сопровождается удешевлением иностранных товаров и удорожанием отечественных. Снижение международной стоимости отечественной валюты приводит к обратной ситуации [73].

На графике можно рассмотреть динамику курса доллара США к рублю (устанавливался Центральным Банком России) – основной международной валюте:

Рисунок 3.10 – Динамика курса доллара США к рублю

Как видно по рисунку 3.10, в конце 2014 – начале 2015 гг. произошел резкий скачок валютного курса. Это привело к удорожанию импортных товаров, а также отечественных товаров, которые производятся на основе импортного сырья. В значительной мере это влияние (совместно с запретом на ввоз импортных товаров) ощутил на себе розничный рынок продуктов питания.

Далее рассматривается доля импорта в количестве продаж анализируемых групп: «Пиво», «Вино» и «Конфеты» (по конкретному магазину розничной сети):

Таблица 3.3

Доля импорта в разрезе товарных групп

Товарная группа	**Доля импорта, %**
Пиво	4,83
Конфеты	8,92
Вино	55,73

Как видно из Таблицы 3.3, максимальная доля импорта в продажах у группы «Вино». Следовательно, в данном сегменте должно было произойти максимальное удорожание и в последствие – падение спроса. Для подтверждения этого факта проводится корреляционный анализ курса доллара и продаж данной товарной группы. Так как влияние курса на покупательский спрос оценивается как отложенное (удорожание товаров происходит не моментально), то необходимо рассматривать корреляционный анализ с лагами.

Рисунок 3.11 – Корреляция валютного курса и продаж группы «Вино»

Исследуя зависимость, можно сказать, что она отрицательная на всех уровнях лага. Т.е. при росте валютного курса продажи вина всегда падают по причине повышенных цен на товар. При этом видно, что максимальная корреляция (по модулю) достигается при лаге в 98 дней, далее происходит спад. Причиной такой разницы и повышенной зависимости именно через 70-100 дней является постепенная реакция розничного рынка на изменение цен. Причины этого освещены ниже:

- Осуществление больших закупочных программ перед ростом цен или сразу после роста цен на товар.
- Наличие остатков товара в достаточном количестве, чтобы осуществлять политику заниженных цен.
- Осуществление политики постепенного роста цен для предотвращения возникновения шоковых ситуаций.
- Расширение акционных программ, направленных на стимуляцию сбыта товара, повышения привлекательности розничной сети в ухудшающихся экономических условиях.

К сожалению, эти причины говорят о том, что для использования показателя валютного курса в оперативном прогнозировании спроса необходимы сложные аппаратные построения. Это связано, прежде всего, с наличием влияния структуры текущих остатков товара на эффект от валютного курса. Это приводит к логической ошибке, когда проводится прогнозирование товарного спроса с включенным в модель уровнем остатков, который может быть управляемой величиной.

Анализ валютного курса на данном этапе используется в экспертных моделях, в особенности, когда перед резким скачком осуществляется закуп товара у поставщика, а затем его продажа с более высоким уровнем наценки.

Анализ влияния праздников

Необходимо рассмотреть влияние праздничных дней на динамику количества чеков и продажи товарных групп. Под праздничными днями понимаются:

- Государственные праздники, во время которых объявляется официальный выходной;
- Предпраздничные дни, в силу специфики розничной торговли: основные объемы продаж происходит в вечернее время после рабочего дня. Поэтому делается предположение, что в предпраздничный день покупательский поток возрастает еще больше, так как происходит основной закуп населения для проведения праздника;
- Постпраздничные дни – т.е. официальные выходные после дня праздника. Их влияние на покупательский спрос неоднозначен и его также необходимо изучить.

Каждый праздник может влиять по-разному как на количество покупателей, так и на состав потребительской корзины. Поэтому исследуется соответствующее графическое представление для подтверждения гипотезы о качественно разном влиянии того или иного праздника. Выделяются в отдельные подгруппы предпраздничные и постпраздничные дни, а также предновогодние.

Рассматривается влияние праздников на потребительскую корзину на примере 3 товарных групп.

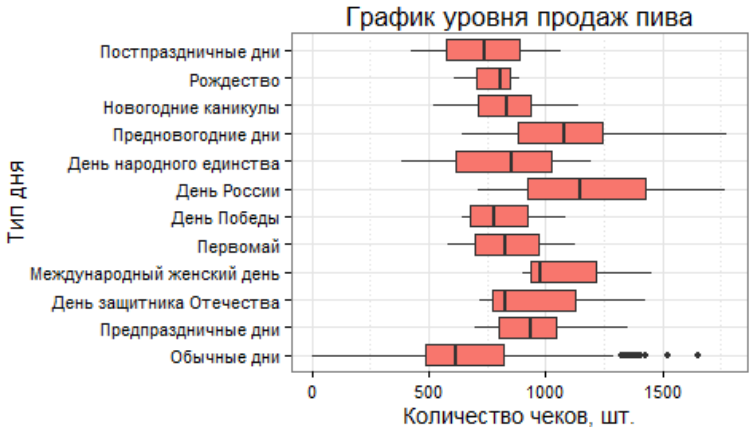


Рисунок 3.12 – Продажи пива в зависимости от типа дня и праздника

По группе «Пиво» наблюдается высокая активность в продажах во время празднования Дня России – среднее количество 1200 проданных штук. При этом распределение продаж в этот день достаточно распластанное. Это может быть связано с таким фактором как разность погоды в день праздника в 2012 – 2015 гг. Подробнее влияние погоды (температурного режима) будет рассмотрено в следующем разделе.

В целом повышенный спрос наблюдается во все дни и варьируется в средних величинах от 700 до 1200 проданных за день штук.

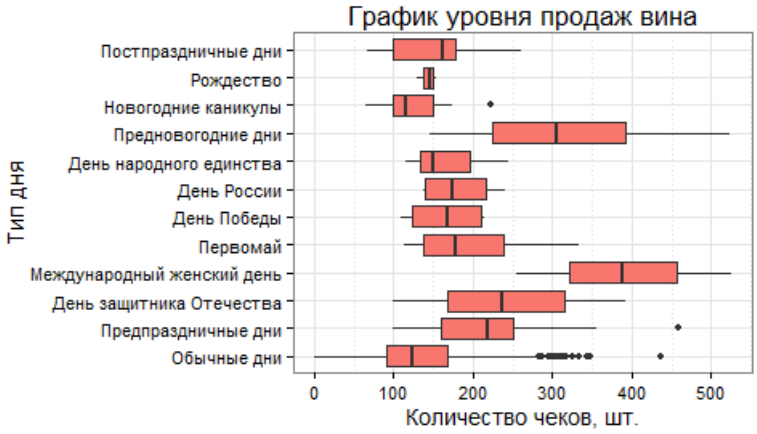


Рисунок 3.13 – Продажи вина в зависимости от типа дня и праздника

Ожидаемо, что повышенные продажи вина наблюдаются в предновогодние дни и в Международный женский день. Средние продажи в эти дни в 3-4 раза выше продаж в обычные (непраздничные) дни. Некоторое снижение покупательской активности по этому напитку наблюдается в новогодние каникулы. Также если рассмотреть на графике тип «Обычные дни», то можно увидеть скопление точек выше третьего квартиля – это могут быть какие-либо неучтенные праздники или дополнительная маркетинговая активность.

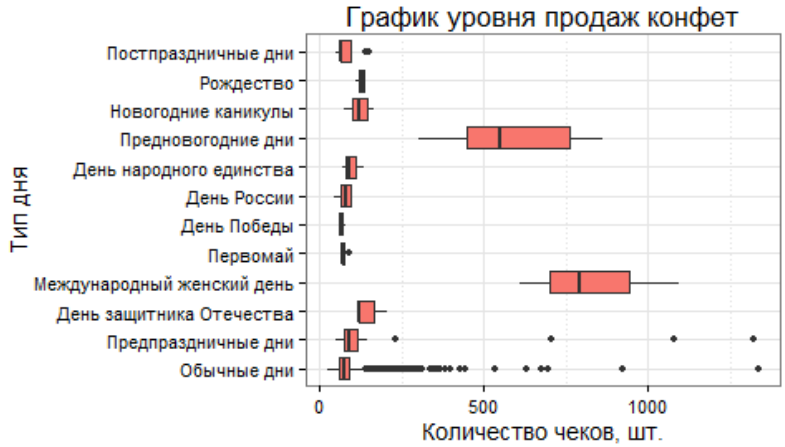


Рисунок 3.14 – Продажи конфет в зависимости от типа дня и праздника

Конфеты обладают средним повышенным спросом в предновогодние дни и Международный женский день. В эти дни конфеты активно покупаются на подарок целевым потребителям – женщинам и детям. Также можно наблюдать большое скопление точек выше третьего квартиля в обычные дни. Данные повышенные продажи могут быть связаны с более ранней реакцией на наступление новогодних праздников – конфеты закупаются покупателем на детские предновогодние праздники, подарки родным и близким и т.д.

Исходя из проделанного выше анализа делается несколько выводов:

- В праздничные дни растет как количество покупателей в магазине, так и покупательский спрос на разные товарные группы;

- Каждый праздничный день влияет на количество покупателей и состав покупок по-разному. 8 марта активно продаются конфеты и вино, а 12 июня – пиво;
- Исходя из степени влияния праздников на покупательский спрос и его структуру необходимо учитывать их при построении прогнозов;
- Помимо государственных праздников существуют также неучтенные религиозные и профессиональные, которые имеют массовый характер. Необходимо также включить в итоговое построение прогнозной модели в качестве дополнительных переменных.

Анализ влияния погоды (температурного режима)

Для понимания наличия возможных дополнительных факторов, влияющих на спрос, необходимо рассмотреть влияние температурных условий на покупательскую активность и продажи товарных групп. Для этого проанализируем температурный ряд с 01.04.2012 по 18.10.2015, который был взят из архива работы метеостанций города Ижевска:

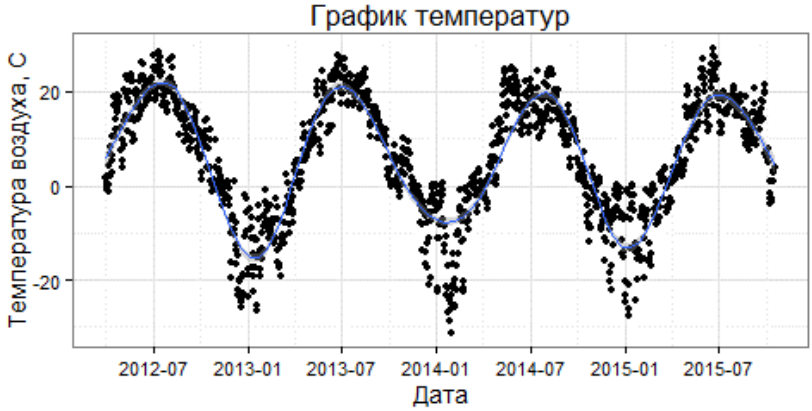


Рисунок 3.15 – График температур

Видно четкую синусоидальную структуру ряда температур. Проанализируем зависимость температуры с показателями: количество чеков, продажи товарных групп «Пиво», «Вино» и «Конфеты».

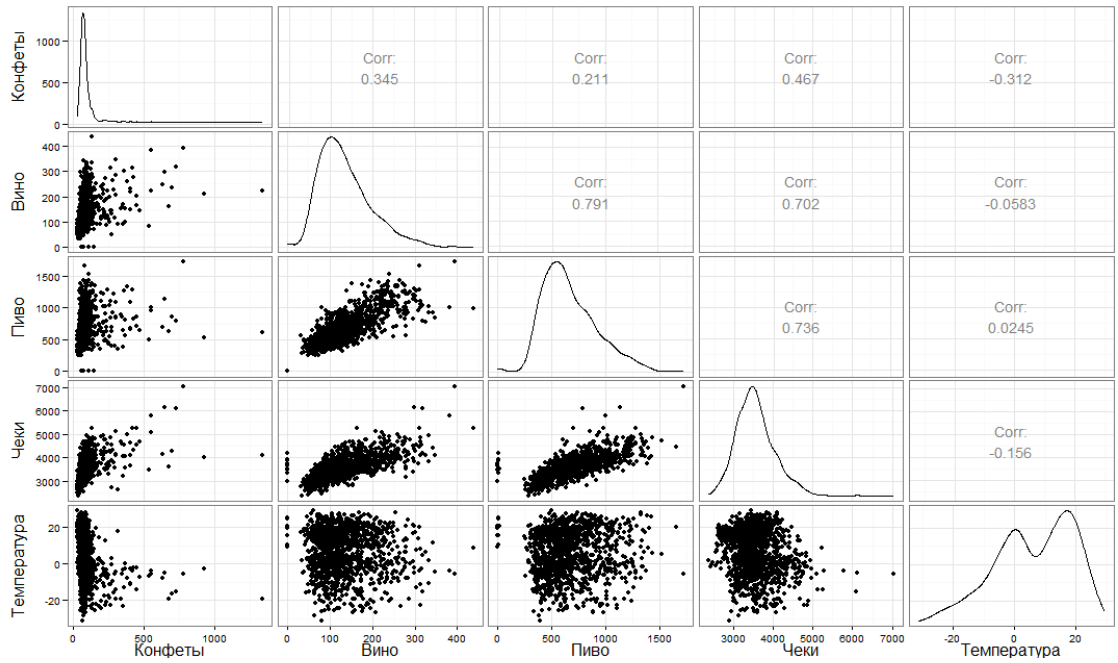


Рисунок 3.16 – График корреляций между температурой и основными показателями

Из анализируемых данных убираются праздничные дни, так как они могут достаточно сильно исказить зависимости. Это касается прежде всего предновогодних дней, когда покупательский спрос вырастает в несколько раз. На графике (рис. 3.16) необходимо исследовать диаграммы рассеяния, графики плотности распределения рассматриваемых величин и уровень корреляции.

Как видно, все показатели имеют слабую линейную взаимосвязь с температурой. Максимальный коэффициент корреляции (по модулю) достигается с группой «Конфеты»: равен -0,321. Тем не менее, при раннем анализе обнаружилось, что данная взаимосвязь может быть обусловлена более ранней реакцией на наступление новогодних праздников, что пересекается с низким температурным режимом.

Для того, чтобы сделать вывод об итоговом влиянии температуры на покупательскую активность, анализируется предположение, что на продажи пива особенно влияет жаркая погода (влияние со знаком «+»). Напротив, в холодную погоду – объемы спроса на данный спиртной напиток падают. Отдельно выведем диаграмму рассеяния пива и температуры:

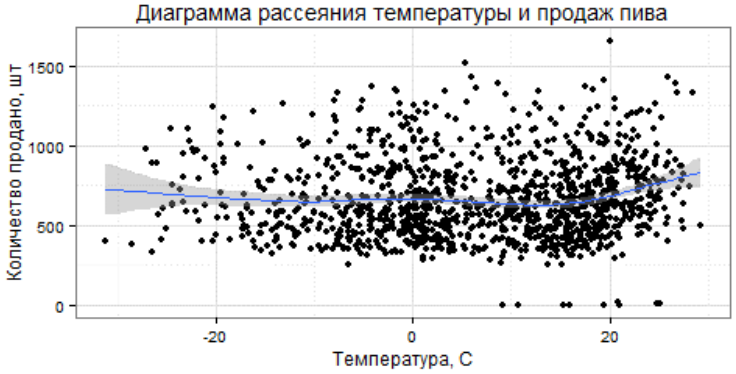


Рисунок 3.17 – Диаграмма рассеяния продаж пива и температур

Как видно корреляция продаж пива и температуры стремится к 0. Тем не менее, можно увидеть, что тренд продаж меняется на уровне от +15 градусов по Цельсию. Эта часть данных исследуется подробнее:

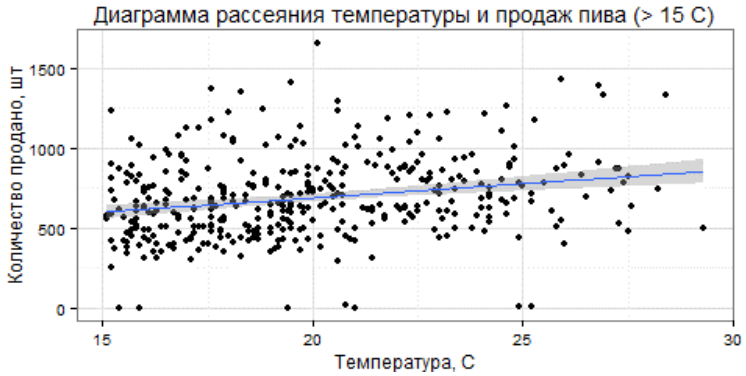


Рисунок 3.18 – Диаграмма рассеяния продаж пива и температур (> 15 C)

В данных условиях, тренд уже более выраженный и возрастающий. Если рассчитать коэффициент корреляции, то он будет равен 0,229, что почти в 10 раз превышает коэффициент корреляции, рассчитанный на весь временной ряд 0,0245 (см. рис. 3.17). Это говорит о более выраженной зависимости между температурой и продажами пива при теплой и жаркой погоде.

Рассматривается также диаграмма рассеяния при температуре меньше -20 C:

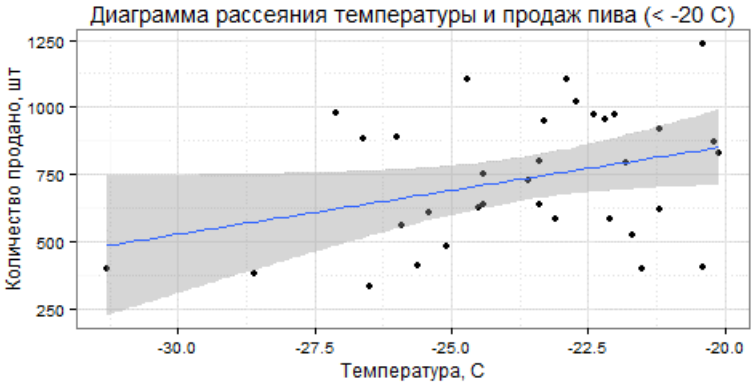


Рисунок 3.19 – Диаграмма рассеяния продаж пива и температур (< -20 C)

На данном графике также видно, что существует более четко выраженный возрастающий тренд, коэффициент корреляции равен 0,342. После оценки зависимостей (рис. 24 – рис. 26), можно отметить, что гипотеза о поведении покупательского спроса на пиво, реагирующего на аномальные температуры, имеет место быть.

Исходя из анализа влияния температуры, можно сделать следующие выводы:

- Нет четкой линейной зависимости между температурным рядом и показателями покупательской активности;
- По некоторым товарным группам существует неоднородная (возможно нелинейная) зависимость продаж от температуры. Например, в теплую погоду с каждым градусом

продажи пива растут быстрее, чем обычно. Еще быстрее они падают, когда наступает действительно холодная погода (ниже 20 С). Подобные предположения можно доказать и выстроить для групп минеральных и газированных напитков, овощей, фруктов и т.п.

Выводы о факторах влияния на покупательский спрос

Исходя из описанного в разделе анализа, делается вывод об обязательном включении следующих факторов в модель прогнозирования покупательского спроса:

1. Внутренняя динамика продаж товара. Каждый товар имеет специфичный покупательский спрос, характерный только для него. В базу любого моделирования необходимо включать динамику значений покупательского спроса с некоторым лагом. Анализ показал, что предыдущие значения продаж влияют на сегодняшние (наличие автокорреляции). Кроме того, так как разработка ведется для прогнозирования спроса на конкретный товар, необходимо ввести в качестве предиктора модели общие продажи по товарной группе;
2. Количество покупателей в магазине. Покупательский поток статистически значимо имеет зависимость с продажами товара. Логично предположить, что этот фактор является первичным – покупатель в розничном магазине далеко не всегда планирует весь спектр покупок. Поэтому фактор будет включен в общую модель прогнозирования спроса.
3. Календарные праздники. Эффект от наличия календарного праздника является неоспоримым. В модель прогнозирования потребительского спроса данная переменная будет включаться как фиктивная.
4. Температурный режим. Корреляционный анализ показал неоднозначность зависимости между температурой и продажами товаров. Тем не менее, при изучении данных на разных температурных диапазонах были выявлены зависимости при достаточно больших и достаточно низких температурах.

После проведения корреляционного и графического анализа информации представляется возможным выделить конкретный состав переменных для прогнозирования спроса. Данный первоначальный состав переменных описывается в разделе 2.2. Методологически он был выбран исходя из представленного в текущем разделе отчета, направленного на то, чтобы подтвердить первоначальные гипотезы о факторах влияния на покупательский спрос.

3.1.2. Реализация эвристического алгоритма подбора переменных

Для отбора переменных используется эвристический алгоритм, основанный на применении базового алгоритма, а также сравнении его оценок с наиболее простым. На этапе алгоритма на рис. 2.3 после определения всех переменных и их размерностей проводится оценка коэффициентов множественной линейной регрессии с аддитивным эффектом (2.2). В таблице частично приведена оценка коэффициентов модели (всего в модели 60 переменных):

Таблица 3.4

Коэффициенты модели линейной регрессии (без нелинейных преобразований)

Показатель	Оценка коэффициента	Ст. ошибка	t-значение	p-значение
Коэффициент β_0	-1,467	0,175	-8,369	0,000
Спрос лаг 1	0,371	0,002	208,677	0,000
…	…	…	…	…
Спрос лаг 7	0,344	0,002	195,087	0,000
Товарный кластер 2	0,127	0,047	2,671	0,008
Товарный кластер 3	0,114	0,051	2,254	0,024
Товарный кластер 4	0,158	0,084	1,880	0,060
…	…	…	…	…
Наличие акции	3,566	0,105	34,061	0,000
Наличие акции лаг 1	-2,638	0,058	-45,763	0,000
Уровень скидки	5,068	0,382	13,255	0,000
Средняя температура	0,001	0,001	1,957	0,050
Номер дня в году	0,000	0,000	0,651	0,515
…	…	…	…	…
Емкость (вес)	-0,139	0,019	-7,250	0,000
Кол-во чеков	0,000	0,000	18,742	0,000
Вторник	0,265	0,028	9,632	0,000
…	…	…	…	…
Воскресенье	0,239	0,029	8,357	0,000
Пасха	-0,384	0,172	-2,227	0,026
14 февраля	-0,319	0,181	-1,769	0,077
23 февраля	-0,012	0,149	-0,081	0,936
…	…	…	…	…
Непраздничный день	-0,348	0,149	-2,334	0,020
Кол-во SKU взаимозам-х по цене	-0,008	0,001	-11,765	0,000
Кол-во SKU взаимозам-х по цене (акция)	-0,015	0,004	-4,309	0,000

Модель имеет объясненный $R^2 = 0{,}7549$, MSE_{lm} на тестовой выборке равен 10,64.

По полученной модели видно, что большинство коэффициентов являются в достаточной степени значимыми (p-значение стремится к 0). Подтверждается первоначальная гипотеза, изложенная в разделе 2.3, о коэффициенте при показателях «Кол-во SKU, взаимозаменяемых по цене» (все и только акционные) - они являются отрицательными и значимыми, поэтому при росте количества взаимозаменяемых SKU спрос на конкретный товар снижается.

Видно, что множественная линейная регрессия без какого-либо нелинейного преобразования не справляется с задачей получения лучшей модели по критерию качества MSE (исходя из алгоритма на рис. 2.3): $MSE_{lm} > MSE_{MA}$, где оценка MSE_{MA} равна 10.55. Для улучшения результата в рамках модели необходимо произвести дальнейшие изменения формулы линейной регрессии. В ходе преобразований также итерационно повторяется эвристический алгоритм подбора переменных.

Итоговая модель интерпретируется следующей формулой:

$$y = \beta_0 + (a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_k y_{t-k}) \times (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{m-l} x_{m-l}) + \cdots + \beta_m x_m + \varepsilon, \quad (3.1)$$

где y - целевая (зависимая) переменная, $y_{t-1}, y_{t-2}, \ldots, y_{t-k}$ – лагированные значения ряда, $a_1, a_2, \ldots, a_k$ – авторегрессионные коэффициенты модели, m – количество независимых переменных, $x_1, x_2, \ldots, x_m$ - независимые переменные и $\beta_1, \beta_2, \ldots, \beta_m$ – коэффициенты при зависимых переменных, рассчитанные методом наименьших квадратов, β_0 – свободный коэффициент модели, ε – случайная ошибка модели.

Видно, что спецификация модели в ходе реализации алгоритма несколько изменилась – определена мультипликативная зависимость между авторегрессионным функционалом и несколькими переменными из исходного набора. Здесь необходимо отметить, что в ходе моделирования было оценено 194 коэффициента, что в 3,23 раза превышает количество переменных в первоначальной версии регрессионной модели. Тем не менее, добавление нелинейных зависимостей между переменными позволило достичь результата: $MSE_{lm} = 8{,}87$, что является меньшим значением чем MSE_{MA}.

Ниже приведена таблица с параметрами оценки некоторых из рассматриваемых переменных:

Таблица 3.5

Коэффициенты модели линейной регрессии (с нелинейными преобразованиями переменных)

Показатель	Оценка к-та	Ст. ошибка	t-значение	p-значение
Коэффициент β_0	3,215	1,515	2,122	0,034
Спрос лаг 1 * Товарный кластер 1	0,310	0,020	15,507	0,000
Спрос лаг 1 * Товарный кластер 2	0,358	0,018	20,468	0,000
Спрос лаг 1 * Товарный кластер 3	0,642	0,034	19,006	0,000
…	…	…	…	…
Спрос лаг 1 * Цена товара	-0,002	0,000	-11,707	0,000
Спрос лаг 1 * Наличие акции	0,410	0,013	30,586	0,000
Спрос лаг 1 * Наличие акции лаг 1	-0,321	0,007	-43,452	0,000
Спрос лаг 1 * вторник	0,098	0,009	11,079	0,000
…	…	…	…	…
Средняя температура	0,002	0,001	3,570	0,000
Кол-во чеков	0,000	0,000	13,919	0,000
Уровень скидки	8,602	0,145	59,420	0,000
Пасха	-4,990	1,539	-3,243	0,001
14 февраля	-4,089	1,544	-2,649	0,008
23 февраля	-4,551	1,521	-2,993	0,003
…	…	…	…	…
Номер дня в празднике	-0,400	0,159	-2,519	0,012
Германия	-0,079	0,029	-2,725	0,006
Иные страны	0,068	0,029	2,342	0,019
Россия	0,438	0,028	15,384	0,000
Чехия	-0,091	0,032	-2,865	0,004
Иные произв-ли	0,270	0,024	11,480	0,000
Произв-ль 1	-0,055	0,026	-2,092	0,036
…	…	…	…	…
Емкость (вес)	-0,027	0,008	-3,545	0,000

Полученная модель имеет объясненный $R^2 = 0{,}8138$, MSE_{lm} на тестовой выборке равен 8,87. Следовательно, в соответствие с заданным алгоритмом выбор исходных переменных и их преобразованных вариантов завершается.

В следующем подразделе описывается процедура получения товарных кластеров, которые используются для текущего результата.

3.1.3. Эффективное разбиение товарных кластеров

Выявление товарных кластеров позволяет решить одну из основных проблем панельных данных – отсутствие индивидуальных меток при моделировании. По сути, кластеры заменяют часть отсутствующей информации и при этом имеют преимущество при группировке отдельных товаров. Исходя из предложенных действий по кластеризации использовались алгоритмы k-средних и EM-алгоритм.

Предварительно был оценен размер *MSE* для линейной модели с нелинейными комбинациями переменных без выделения кластеров, который равен 9,369712. По итогам реализации компьютерных экспериментов по выделению кластеров по двум алгоритмам, результат представлен в таблице 3.6.

Таблица 3.6

Значение MSE на тестовой выборке при реализации кластерного анализа

Вид алгоритма	Количество кластеров			
	2	5	10	15
K-means	8,88359	8,93621	8,94339	8,95204
EM-алгоритм	8,97449	8,87694	9,96271	8,87094

Видно, что реализация алгоритма k-средних ухудшается с ростом количества кластеров, в то время как EM-алгоритм выделения кластеров работает лучше при их увеличении. Кроме того, качество прогнозирования увеличивается сильнее именно при EM-алгоритме, что позволило сделать вывод об адекватном использовании его при реализации основной задачи. В качестве готового решения использовался EM-алгоритм с 15 кластерами. На основе данного разбиения кластеров, был выведен минимальный *MSE* равный 8,87094, что на 5,3% лучше, чем модель без использования кластерного разбиения. Использование большего количества кластеров могло привести к неравномерной структуре по количеству объектов в кластере. Следовательно, результат работы алгоритмов мог быть неустойчив.

3.2. Реализация прогнозирования временных рядов

Прогнозное моделирование производилось для трех ключевых переменных: количество чеков, количество проданных товаров в группе и температурный режим. В качестве исходных

данных для моделирования количества чеков использовались ежедневные значения в одном из магазинов розничной сети ООО «Гастроном» в период с 01.05.2009 по 30.09.2016. Для проверки настройки параметров и проверки качества модели исходный ряд был разделен на обучение и тест. В качестве дополнительных предикторов используются:

- Праздники государственные и религиозные, а также предпраздничные дни. Кодировались как фиктивные бинарные переменные: $H_l = 0$, если нет события l, $H_l = 1$, если праздник в эту дату есть.
- Номер дня в рамках одного праздничного периода.
- Разложенные ряды Фурье относительно исходного ряда F_k. Подбор размера разложения осуществлялся с помощью параметра k, при этом минимизировалась целевая метрика. Итоговая модель имеет в себе $k = 7$ слагаемых.
- В качестве учета эффекта выплаты заработной платы использовался полином следующего вида:

$$\nu_0 + \nu_1 \cdot cal^5 + \nu_2 \cdot cal^4 + \nu_3 \cdot cal^3 + \nu_4 \cdot cal^2 + \nu_5 \cdot cal, \qquad (3.2)$$

где cal – порядковый номер дня месяца, ν_0 – смещение, которое учитывается непосредственно в модели, $\nu_1, \nu_2, \dots, \nu_5$ – коэффициенты, подогнанные в зависимости от используемого метода. Характер полинома также подбирался экспериментально, с целью снизить ошибку моделирования.

Ниже приведен график прогнозов на тестовой выборке по классическим методам (ARIMA, ETS) прогнозирования и фактические значения по количеству чеков (рис 3.20).

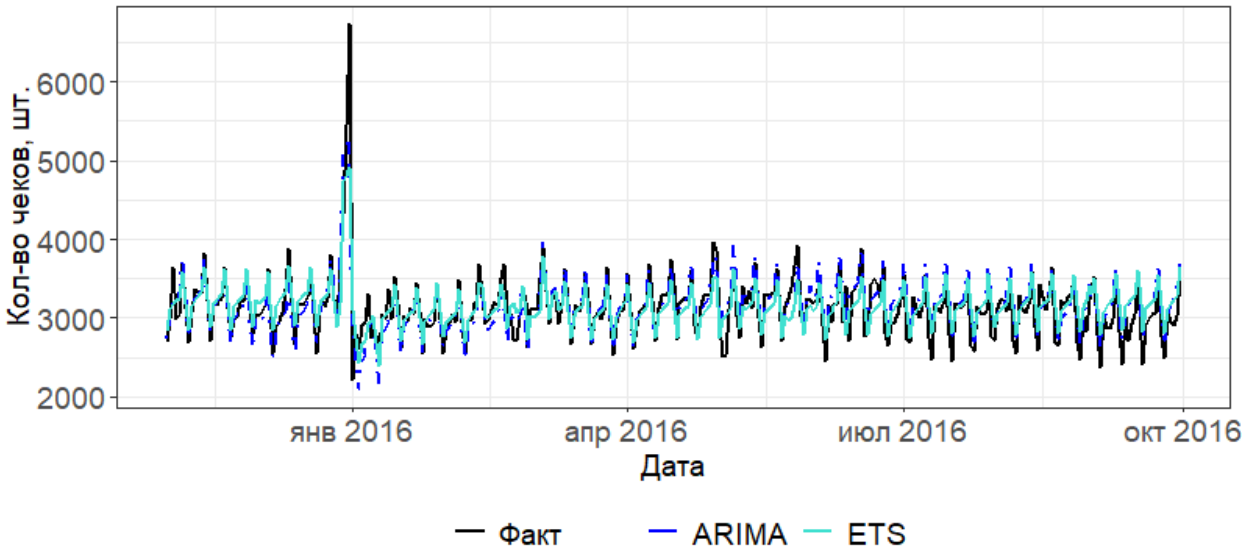


Рисунок 3.20 – Динамика количества чеков с прогнозами (классические методы прогнозирования)

Рисунок 3.20 демонстрирует хорошую аппроксимацию фактических значений классическими методами. Далее приводится аналогичный график с прогнозами чеков с помощью более современных методов (Prophet, CES, BSTS) на рис. 3.21.

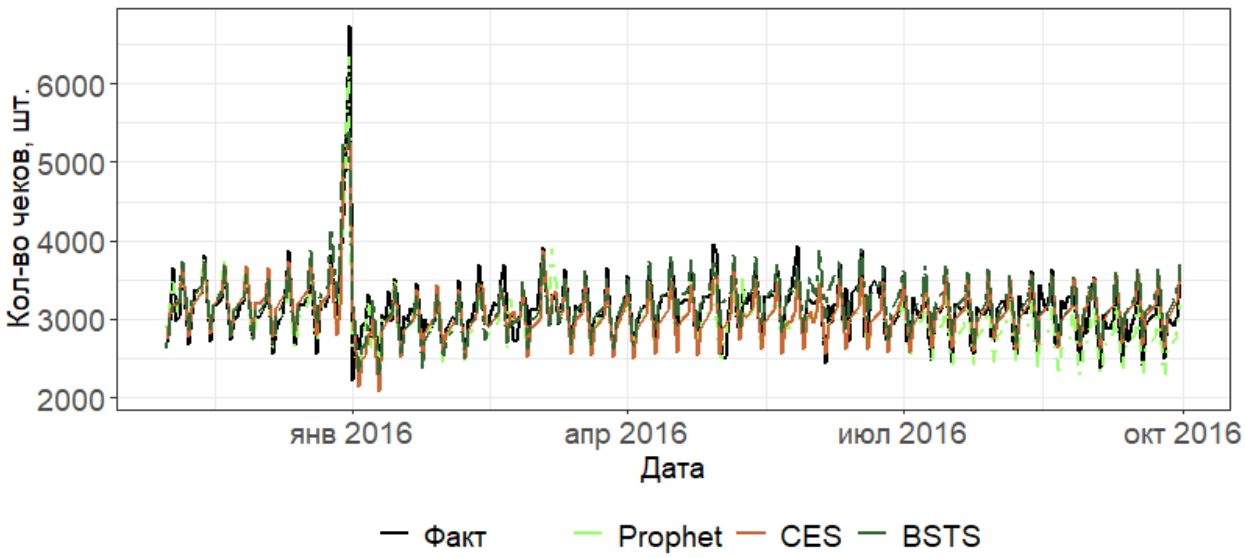


Рисунок 3.21 – Динамика количества чеков с прогнозами (современные методы прогнозирования)

Согласно данным рисунка 3.21, обнаруживается высокая степень подгонки моделей на тестовой выборке.

Средний прогноз формировался по всем методам кроме модели экспоненциального сглаживания (ETS). Такой подход применялся из-за высокой коррелированности прогнозов по методам ETS и CES. Результат такого подхода виден на рис. 3.22.

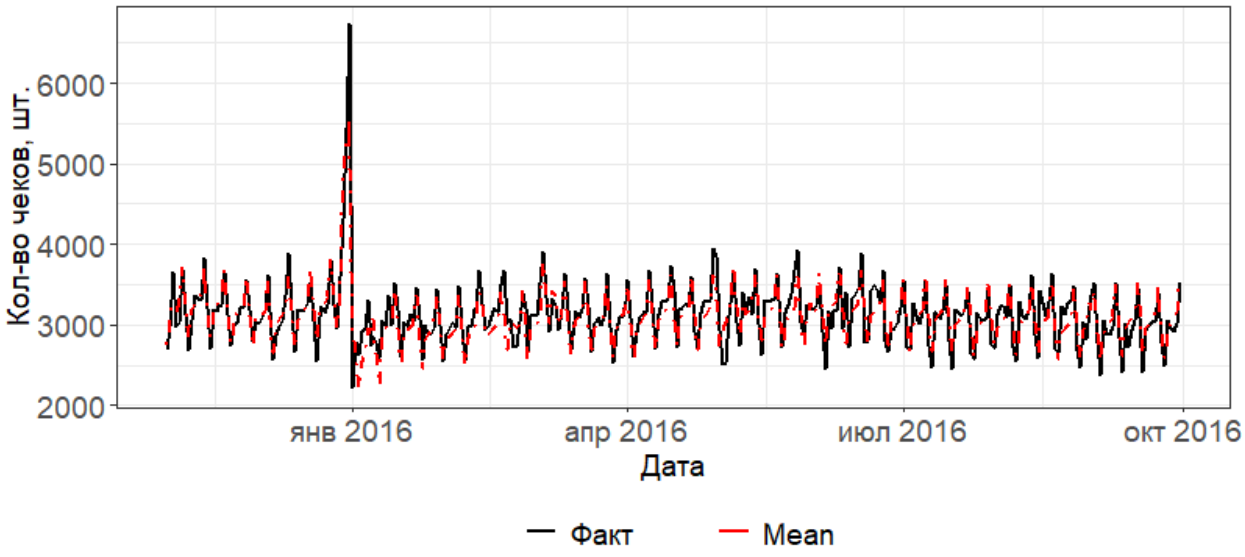


Рисунок 3.22 – Динамика количества чеков со средним прогнозом

По указанным метрикам лучшей из двух моделей является CES. В таблице 3.7 приведены метрики прогнозов.

Таблица 3.7

Результат тестирования методов прогнозирования для количества чеков

Метрика качества	**Prophet**	**ARIMA**	**ETS**	**CES**	**BSTS**	**Mean**
RMSE	240,46	224,40	228,54	227,54	230,86	188,69
MAPE, %	6,05	5,26	5,30	5,40	5,47	4,43

Из таблицы 3.7 видно, что значение среднего арифметического по прогнозам имеет лучшие метрики в сравнении с остальными прогнозами.

Прогнозирование количества проданных товаров осуществлялось с помощью данных по продажам одной из товарных групп розничной точки сети «Гастроном». Период выборки: 01.05.2009 по 30.09.2016. Временной ряд разбивался на обучающий – с 01.05.2009 по 31.12.2015 – и на тестовый – с 01.01.2016 по 30.09.2016. Это позволило обеспечить проверку выстраиваемых моделей.

В качестве основных предикторов выбираются:

- Количество чеков или характеристика покупательского потока. Здесь подтвердилась гипотеза о зависимости данных временных рядов друг от друга.
- Температурный ряд. Также имеет влияние на рассматриваемую группу товаров повседневного спроса.
- Фиктивные переменные для календарных и прочих праздников: $H_l = 0$, если нет события l, $H_l = 1$, если праздник в эту дату есть.
- Номер дня в рамках одного праздничного периода.
- Разложенные ряды Фурье относительно исходного ряда – F_k. Для моделирования данного ряда путем подбора определено $k = 3$ слагаемых.

Выводятся полученные прогнозы для тестовой выборке по классическим методам прогнозирования временных рядов ARIMA и ETS на рис. 3.23.

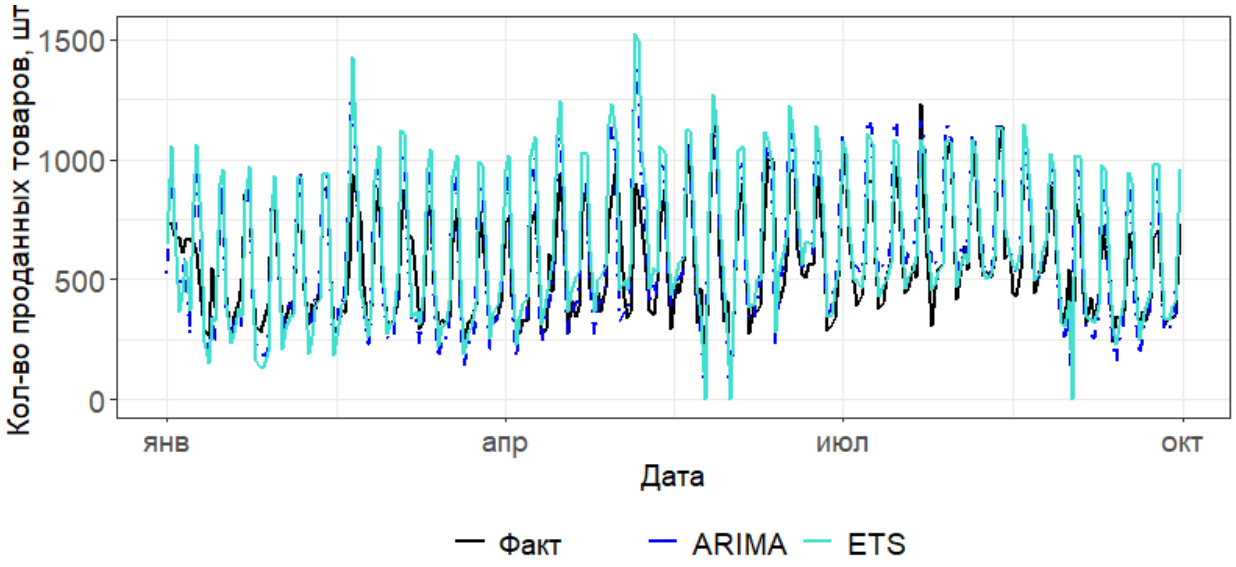


Рисунок 3.23 – Динамика продаж товарной группы с прогнозами (классические методы прогнозирования)

Визуально видна достаточно высокая степень аппроксимации результатов. Далее следует рассмотреть прогнозирование современными реализациями прогнозирования временных рядов на рис. 3.24.

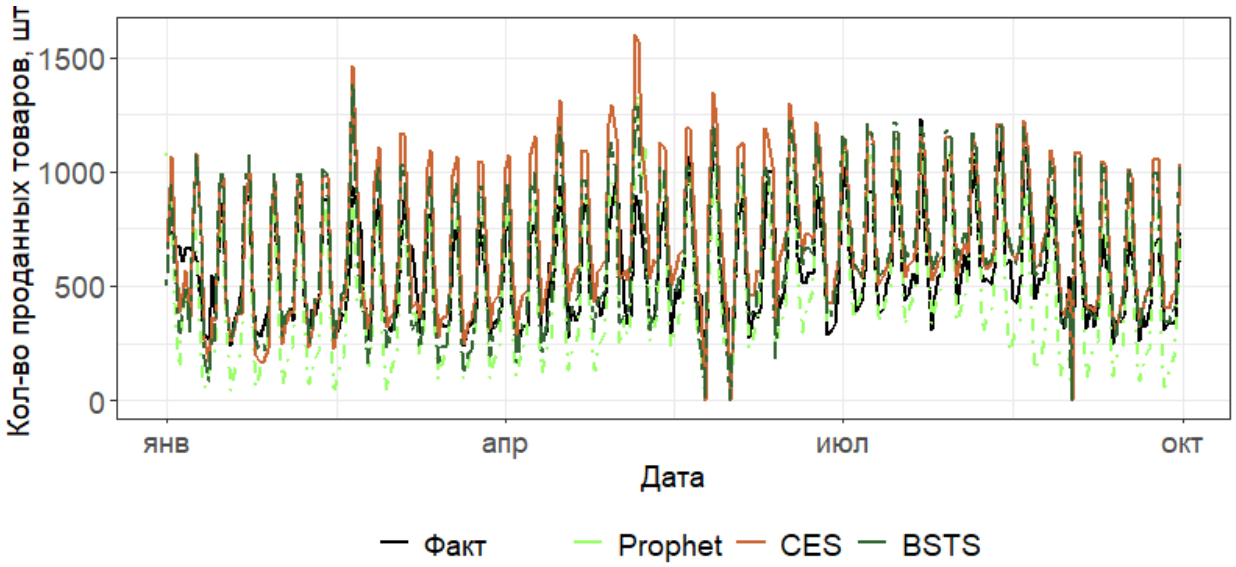


Рисунок 3.24 – Динамика продаж товарной группы с прогнозами (современные методы прогнозирования)

В среднем, колебания по результатам современных методов прогнозирования существенно выше и приводят к большей ошибке на тестовой выборке. По итогам анализа метрик, можно сказать, что итоговым прогнозирования является ARIMA.

Далее выводится итоговый прогноз на рис. 3.25.

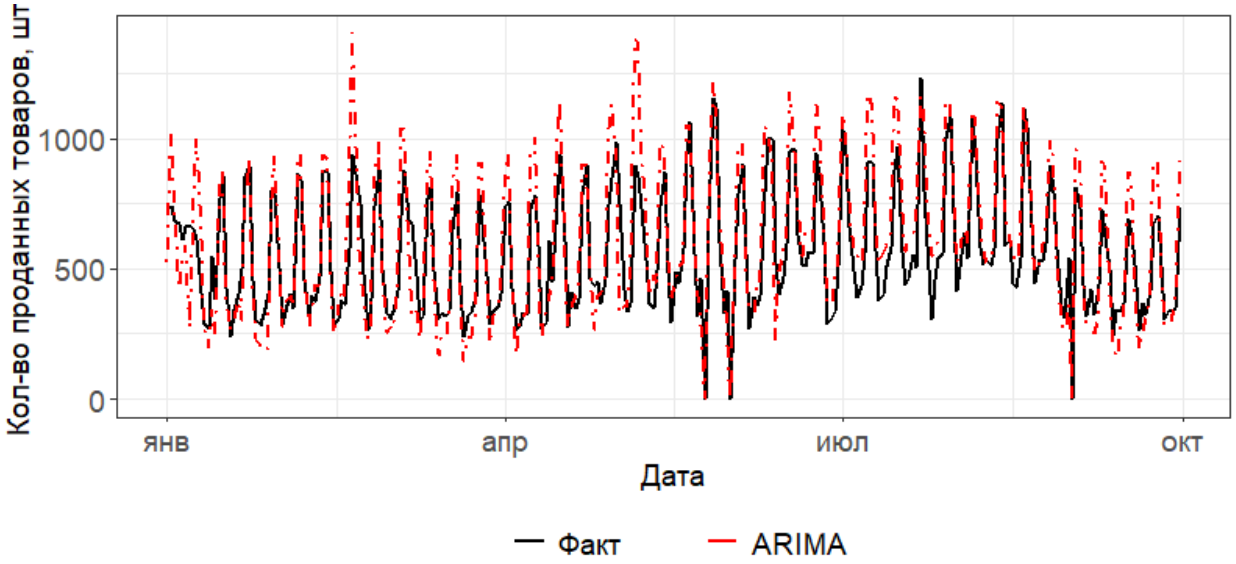


Рисунок 3.25 – Динамика продаж товарной группы с итоговым прогнозом

Визуально, итоговый прогноз продаж допускал некоторые возможности серьезных отклонений. Это связано с оценкой праздничных дней. На практике необходимо делать экспертную корректировку прогнозов в праздники, чтобы снизить ошибку прогнозирования.

Основные метрики прогнозов представлены в таблице 3.8.

Таблица 3.8

Результат тестирования методов прогнозирования для прогноза продаж

Метрика качества	Prophet	ARIMA	ETS	CES	BSTS	Mean (ARIMA, ETS)
$RMSE$	170,21	136,66	156,00	195,06	160,37	142,96
$MAPE$, %	28,93	18,52	20,40	28,59	22,89	18,76

По оценке основной метрики $RMSE$ комбинированный прогноз из лучших методов менее оптимален, поэтому было принято решение об использовании ARIMA в качестве итогового. Точечные экспертные корректировки в праздничные дни и при реализации сложных маркетинговых мероприятий позволят улучшить качество прогноза.

Для построения прогнозов температурного режима использовалась выборка ежедневных средних температур в районе города Ижевска на основании данных, полученных от метеостанций с 01.01.2008 по 30.09.2016. Также, общая выборка делилась на обучение и тест: с 01.01.2008 по 30.10.2015 и с 01.11.2015 по 30.09.2016 соответственно. В качестве данных для моделирования использовался только сам временной ряд температур, иные предикторы не рассматривались. Ниже приведен график (рис. 3.26) прогнозов на тестовой выборке по трем методам прогнозирования, средний прогноз по двум лучшим методам и фактические значения температур.

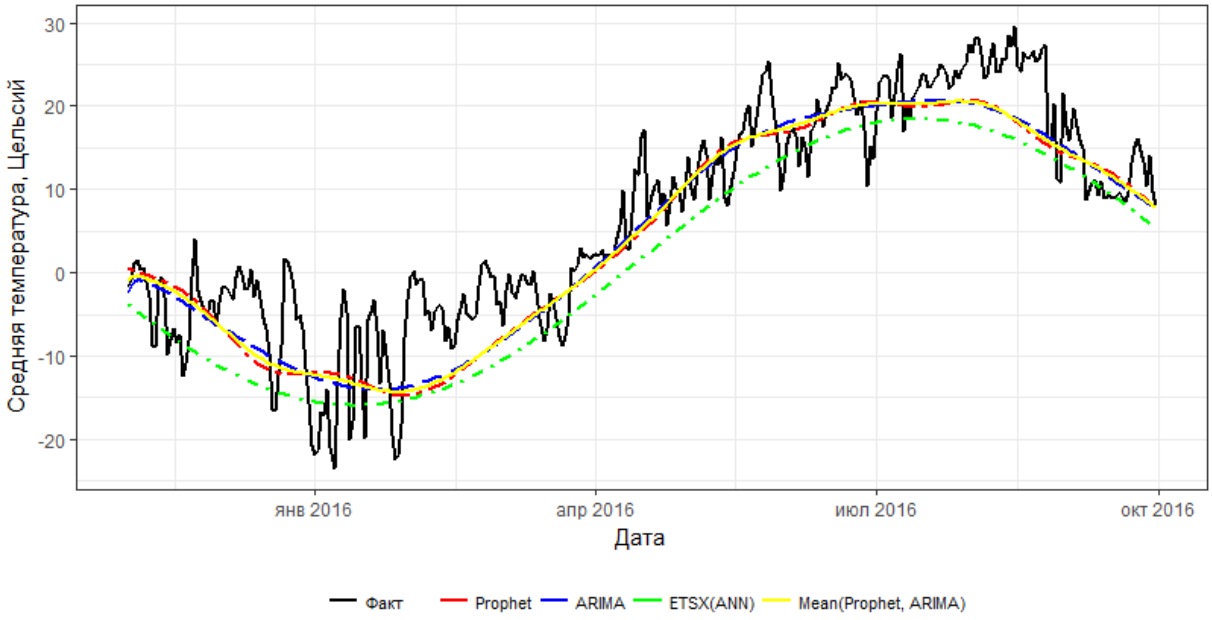


Рисунок 3.26 – Динамика средних температур с прогнозами

Выбор только трех методов прогнозирования из представленных, обусловлен избыточностью применения более продвинутых методов прогнозирования погодных условий при малом объеме данных. При этом, более серьезный подход для данной задачи ресурсозатратен, так как решается проблема управления торговым предприятием.

Приведем таблицу 3.9 с основными метриками прогнозов.

Таблица 3.9

Результат тестирования методов прогнозирования для температурного режима

Метрика качества	Prophet	ARIMA	ETS	Mean(Prophet, ARIMA)
$RMSE$	5,60	5,36	6,85	5,46
$MAPE$, %	201,52	195,49	275,41	197,95

Таблица 3.9 показывает значимость выбора метрики для выявления наиболее эффективного алгоритма прогнозирования. Метрика $MAPE$ показывает себя не лучшим образом, так как исходный временной ряд имеет как положительные, так и отрицательные значения. В случае расчета $MAPE$, если разность $\hat{y}_t - y_t$ делится на фактическое значение $y_t \to 0$, то даже при не столь значительных отклонениях получаем возрастание отношения $(\hat{y}_t - y_t)/y_t$. Поэтому при оценке качества модели следует применять $RMSE$, который дал более четкую картину реализации метода.

Следует также отметить, что для всех моделей в качестве дополнительных регрессоров использовались суммы гармоник, что повысило точность моделирования. Кроме того, наиболее «успешной» моделью является ARIMA, реализованная на базе алгоритма Хиндмана-Хандакара. При этом результат полученный одним алгоритмом является более эффективным, чем комбинация.

3.3. Реализация прогнозирования товарного спроса

В качестве апробации основной методологии прогнозирования спроса выбирались данные для одной товарной группы торговой точки сети ООО «Гастроном». Согласно иерархии прогнозирования, указанной на рис 2.5, модель выстраивается на панельных данных, т.е. присутствуют данные по всем товарам в группе во временном промежутке. Данные сформированы с 01.01.2013 по 30.09.2016. Для формирования обучающей выборки использовалась технология out-of-time – данные формируются исходя из метки времени, что логично, так как по сути моделируются многомерные временные ряды. Обучение – на периоде с 01.01.2013 по 31.01.2016, тестирование результатов проходило на периоде с 01.02.2016 по 30.09.2016. Соблюдалась пропорция разделения данных 70 на 30. Структура данных для обучения моделей следующая:

- Для оценки вероятности ненулевого спроса $P(y \neq 0)$ использовался весь объем данных.
- Для построения регрессионной оценки $\widehat{D}$ данные фильтровались по признаку $y \neq 0$, что позволило моделировать только правую часть мультимодального распределения спроса, скошенного к нулевым значениям.

– Итоговая оценка $\hat{y}$ производилась на всем объеме тестируемых данных как для нулевых, так и ненулевых значений.

Реализация некоторых методов прогнозирования была ограничена техническими возможностями. Это является разумным ограничением, так как практическая реализация должна иметь превалирующее значение в значении практического применения методологии.

3.3.1. Оценка вероятности ненулевого спроса

Согласно выстроенной методологии прогнозирования спроса первым этапом является оценка вероятности ненулевого спроса. Для этого применялись указанные методы машинного обучения.

На исходных данных оценивалась модель логистической регрессии:

$$p(X) = \frac{e^{BX}}{1 + e^{BX}}, \tag{3.3}$$

где BX можно точно раскрыть как

$$BX = \beta_0 + (a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + a_5 y_{t-5} + a_6 y_{t-6}) \times B_1 C + B_2 X_2 \tag{3.4}$$

Здесь, β_0 – смещение, a_1, a_2, a_3, a_5, a_6 и $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-5}, y_{t-6}$ – коэффициенты и значение лагов спроса соответственно, B_1 – набор коэффициентов и C – фиктивный набор переменных, которые отражают принадлежность объекта прогнозирования к товарному кластеру, X_2 – набор прочих предикторов с матрицей коэффициентов B_2. Полный набор предикторов, использованных в моделировании представлен в Приложении А. Спецификация подобрана согласно эвристическому алгоритму подбора предикторов (раздел 2.2) с учетом того, что оптимизируется метрика AUC.

Полученные R-кривые и метрика качества AUC представлены на рисунке 3.27.

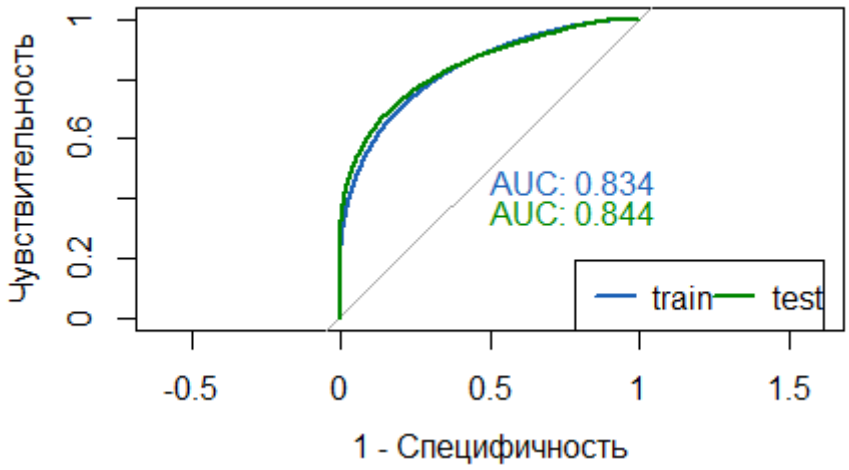


Рисунок 3.27 - ROC-кривые для логистической регрессии

Как видно, качество модели без применения регуляризации, но с отбором переменных характеризовалось как хорошее. Особенностью результатов является значение AUC на тестовой

выборке выше, чем на обучающей. Это может быть связанно с сезонностью в розничной торговле.

После получения результатов по классическому линейному алгоритму применялась регуляризацию. Для этого выбиралась спецификация с чуть более полным набором предикторов:

$$BX = \beta_0 + (a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + a_4 y_{t-4} + a_5 y_{t-5} + a_6 y_{t-6}) \times B_1 C + B_2 X_2 + B_3 X_3, \quad (3.5)$$

где X_3 – набор дополнительно включаемых предикторов для модели с регуляризацией. В формуле (3.5) также добавляется дополнительный лаг спроса y_{t-4}.

Подбор гиперпараметров регуляризации осуществлялся с помощью решетчатого поиска с десятиблочной перекрестной проверкой. Сетка параметров имела следующую структуру:

$$\alpha \in \{0, 0.25, 0.5, 0.75, 1\} \quad \lambda \in \{0, \dots, 10^i, \dots, 10^n\} \quad (3.6)$$

В полученную сетку включено 250 наборов сочетаний гиперпараметров. После проведенного решетчатого поиска оптимальными гиперпараметрами оказались $\alpha = 0$ и $\lambda = 0.01$. Параметр $\alpha = 0$ соответствует ридж-регрессии и штрафу l_2. Далее модель рассчитывалась на всей обучающей выборке.

Итоговый результат виден на рис. 3.28.

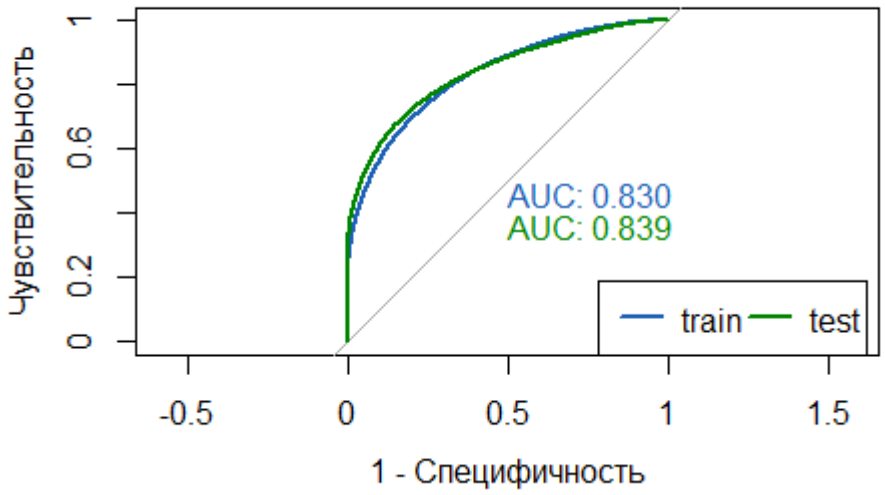


Рисунок 3.28 – ROC-кривые для логистической регрессии с регуляризацией

В данном случае, регуляризация не помогла улучшить результат логистической регрессии, тем не менее, в дальнейшем описан анализ возможности ее использования в итоговой комбинированной оценке $p(X)$.

Далее строился случайный лес классификации. Так как алгоритм являлся достаточно ресурсозатратным, то для его поднастройки использовалась только часть данных. Для этого рассматривался участок данных с самым вариативным спросом, а именно весь спрос в субботние периоды, что представлено в Таблице 3.10.

Таблица 3.10

Статистики товарного спроса по дням недели

День недели	Средний спрос	Стандартное отклонение	Коэффициент вариации
понедельник	1.285992	4.406138	3.426256
вторник	1.502136	5.127294	3.413334
среда	1.663614	5.868962	3.527839
четверг	1.793095	6.330222	3.530332
пятница	3.083569	10.558170	3.424009
суббота	*3.094860*	*11.243541*	*3.632973*
воскресенье	1.995895	6.891492	3.452833

Тестовая (проверочная) выборка для этого кадра данных также была определена строго на данных субботнего спроса.

Сначала составлялся набор возможных вариаций для параметра m – количества случайно отбираемых признаков:

$$m \in \{3, 5, 10, 15\} \tag{3.7}$$

Набор параметров (3.7) полностью соответствовал рекомендациям Лео Бреймана, которые отражены в Таблице 2.3, исходя из того, что предикторов для построения модели использовано $M = 24$.

Производился поиск оптимального параметра с помощью пятиблочной перекрестной проверки с пятью повторами (то есть блоки пять раз формировались заново для одного и того же параметра m). Результат кросс-валидационной проверки параметра m представлен на рисунке 3.29.

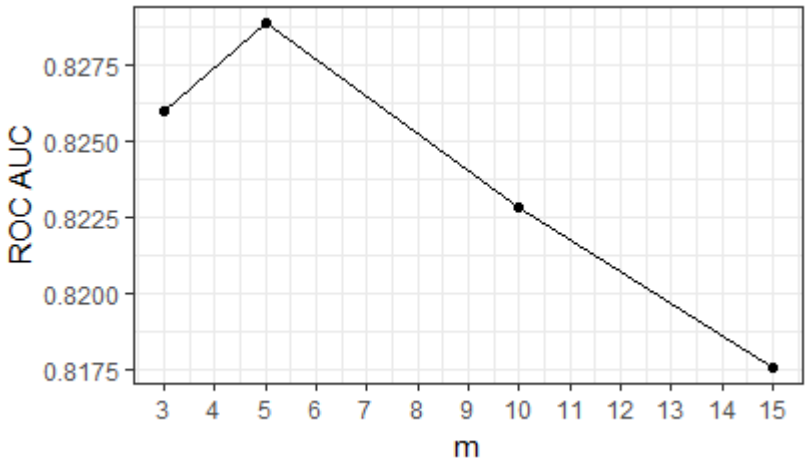


Рисунок 3.29 – Подбор параметра m с помощью перекрестной проверки

Видно, что оптимальное значение метрики $ROC\ AUC = 0.8288462$ достигнуты при $m = 5$.

Далее проводилась регулировка параметра количества деревьев N и расчет AUC для тестовой выборки (также мини-тестовая выборка сформирована на субботнем спросе), что видно на таблице 3.11.

Таблица 3.11

Настройка параметра количества деревьев на тестовой выборке малого размера

N	$ROC\ AUC$
300	0.8249579
500	0.8266745
800	0.8258735
1200	0.8262851

Выводом из таблицы является, что количество деревьев $N = 500$, которое стояло по умолчанию при реализации алгоритма, является оптимальным значением и увеличивает точность результата на тестовой выборке в сравнении с остальными вариациями параметра.

Производился итоговый расчет при $m = 5$ и $N = 500$ для всей обучающей и тестовой выборки. При этом, значение по умолчанию для минимального количества объектов в узлах формируемых деревьев $MN = 10$. Результат в виде ROC-кривых приведен на рисунке 3.30.

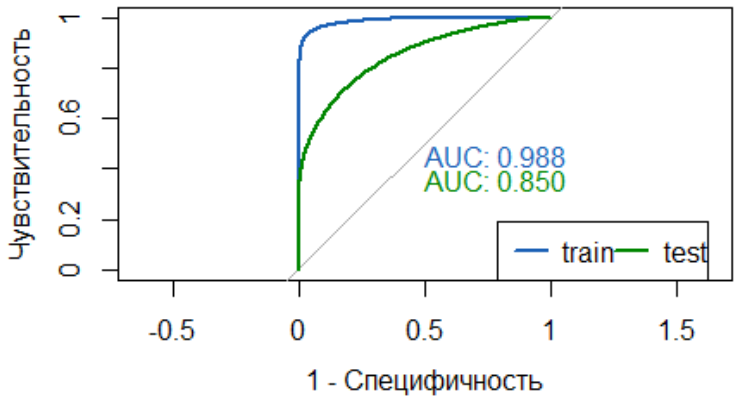


Рисунок 3.30 – ROC-кривые для случайного леса

Видно, что модель довольно переобучена и на обучающей выборке получала оценку значительно выше, чем на тестовой. Тем не менее, уже эта полученная модель случайного леса дает результат немного лучше, чем у логистической регрессии: $AUC = 0.850$ против $AUC = 0.844$. Из-за эффекта переобучения была предпринята попытка упростить модель, увеличивая количество объектов в узле, тем самым упрощая деревья, которые легли в основу вероятностного случайного леса. В ходе экспериментов было выявлено, что можно добиться снижения эффекта переобучения при увеличении MN, что наглядно отражено на рис. 3.31.

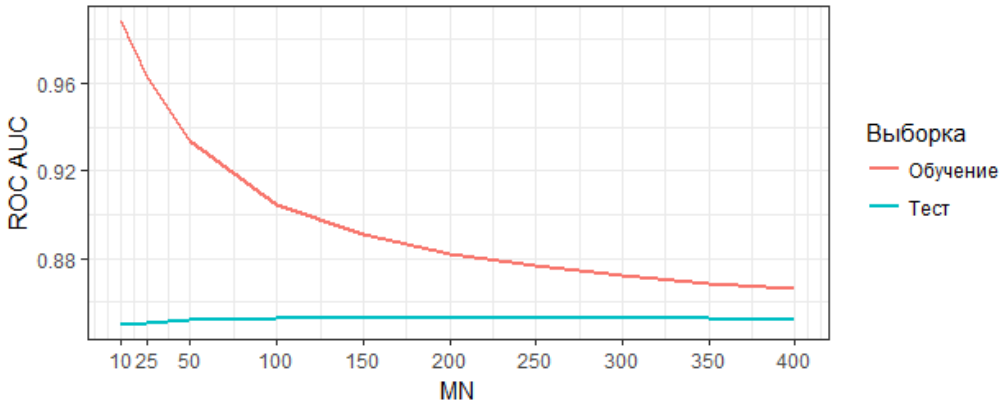


Рисунок 3.31 – Снижение эффекта переобучения

Максимум оценки AUC на тестовой выборке при этом достигалась на $MN = 200 – AUC = 0.853$ (рис. 3.32).

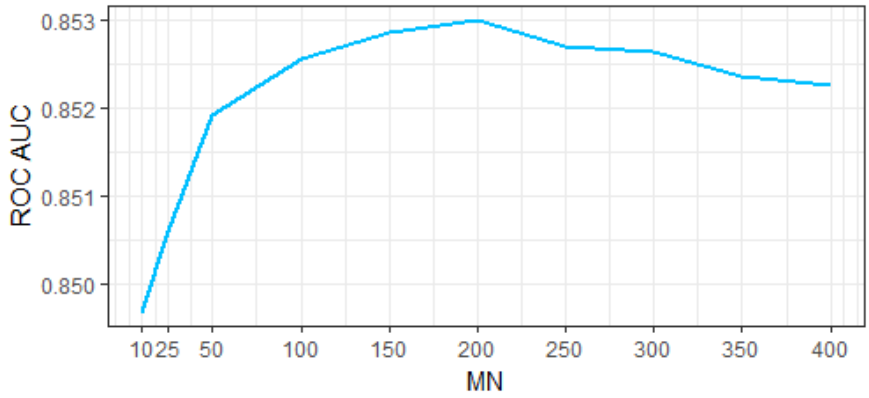


Рисунок 3.32 – Значение AUC в зависимости от количества объектов в узле на тестовой выборке

Следовательно, данные настройки модели случайного леса принимаются за оптимальные. На рис 3.33 выводятся ROC-кривые для анализа итогового результата.

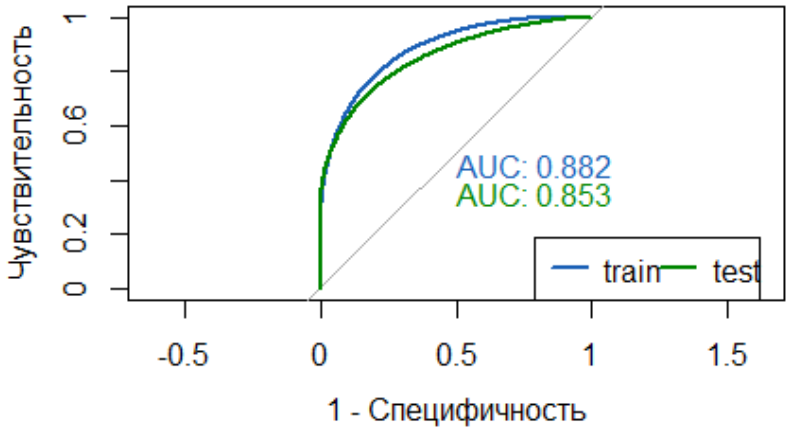


Рисунок 3.33 – ROC-кривые для случайного леса с итоговыми параметрами

Видно, что эффект переобучения значительно снизился – ROC-кривые на обучающей и тестовой выборке значительно приближены друг к другу.

Для реализации градиентного бустинга используется библиотека XGBoost. В качестве основного параметра B выбирались решающие деревья. Метод построения бустинга для классификации был реализован простым перебором сетки параметров из-за высоких затрат на экспериментальный расчет. Для всех описанных в разделе 2.5.3 параметров была составлена следующая сетка:

$$\begin{gathered}\eta \in \{0.02, 0.04, 0.06, 0.08, 0.1\} \\ N \in \{150, 300\} \\ MD \in \{4, 6, 8\} \\ MC \in \{1,3,5\} \\ \delta \in \{0.5, 1\} \\ c \in \{0.5, 0.7, 0.9\}\end{gathered} \tag{3.8}$$

Всего решетка состояла из 540 сочетаний параметров. Модель настраивалась с помощью десятиблочной перекрестной проверки без использования повторений. Результат виден на рис. 3.34.

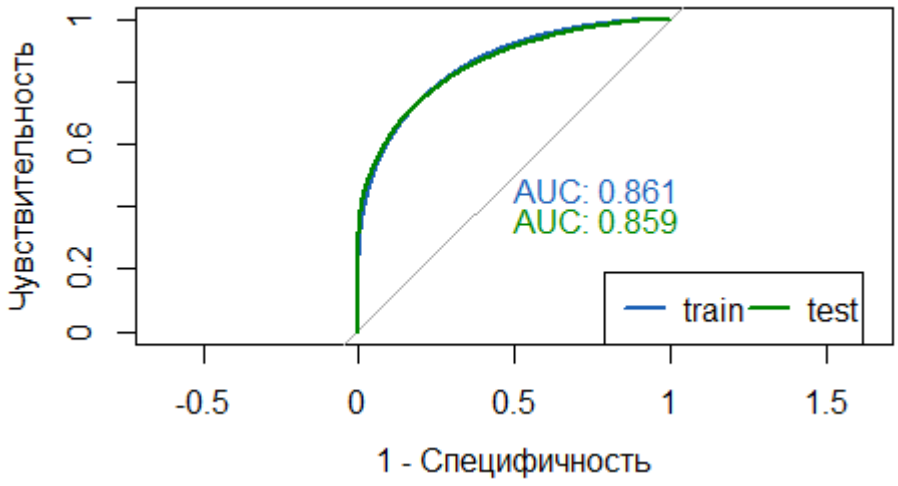


Рисунок 3.34 – ROC-кривые для градиентного бустинга

Результат оценки AUC являлся несколько лучшим, чем у случайного леса: $AUC = 0.859$ вместо $AUC = 0.853$. Кроме того, видно, что переобучение модели имеет практически незаметный эффект, поэтому модель градиентного бустинга является более устойчивой к изменениям в данных. Следовательно, для оценки вероятности ненулевого спроса подобраны все алгоритмы и оценены все результаты.

3.3.2. Решение регрессионной задачи

В виду использования для расчетов регрессионной задачи меньшего по объему набора данных, только для $D \neq 0$, расчеты производятся более эффективно.

Для оценки линейной регрессии использовалась несколько отличная от логистической регрессии структура модели:

$$D = \beta_0 + (a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + a_5 y_{t-5} + a_6 y_{t-6}) \times (B_1 C + B_{21} X_2) + B_{22} X_2 + B_3 X_3, \quad (3.9)$$

где X_2 – часть предикторов, связанных с целевым значением мультипликативно, X_3 – независимые переменные, связанные со спросом аддитивно. Как и ранее, для отбора переменных использовался эвристический алгоритм из раздела 2.2. Простая подгонка модели методом наименьших квадратов без процедуры регуляризации привела к следующему результату на обучающей и тестовой (только для $D \neq 0$) выборках:

$$MSE_{tr} = 27.21967, \quad MSE_{ts} = 23.73885$$
$$MAE_{tr} = 0, \quad MAE_{ts} = -0.09650419, \quad (3.10)$$

где tr – метрика на обучающей выборке, ts – метрика на тестовой выборке.

Применение регуляризации типа (2.37) с незначительными изменениями в количестве предикторов – добавление лага y_{t-4} и некоторых преобразованных переменных – позволило несколько улучшить результат моделирования. В качестве сетки параметров используется (3.6). В качестве перекрестной проверки выбиралась десятиблочная структура. Из всей сетки параметров оптимальными являлись $\alpha = 0{,}25$ и $\lambda = 0.01$. Для указанных параметров регуляризации, результат модели показал следующие оценки ключевых метрик:

$$MSE_{tr} = 27.07184, \quad MSE_{ts} = 23.50026$$
$$MAE_{tr} = 0, \quad MAE_{ts} = -0.1306996 \quad (3.11)$$

Видны небольшие улучшения в метрике MSE и небольшое ухудшение в MAE. В целом результаты (3.10) и (3.11) очень схожи.

Проводились также расчеты для регрессионного случайного леса. В данном случае производился перебор сетки параметров на десятиблочной перекрестной проверке без повторов:

$$N = 500$$
$$m \in \{3, 5, 12, 18, 25, 35\} \quad (3.12)$$
$$MN \in \{10, 25, 50, 75\}$$

Результат кросс-валидационной проверки представлен на рисунке 3.35.

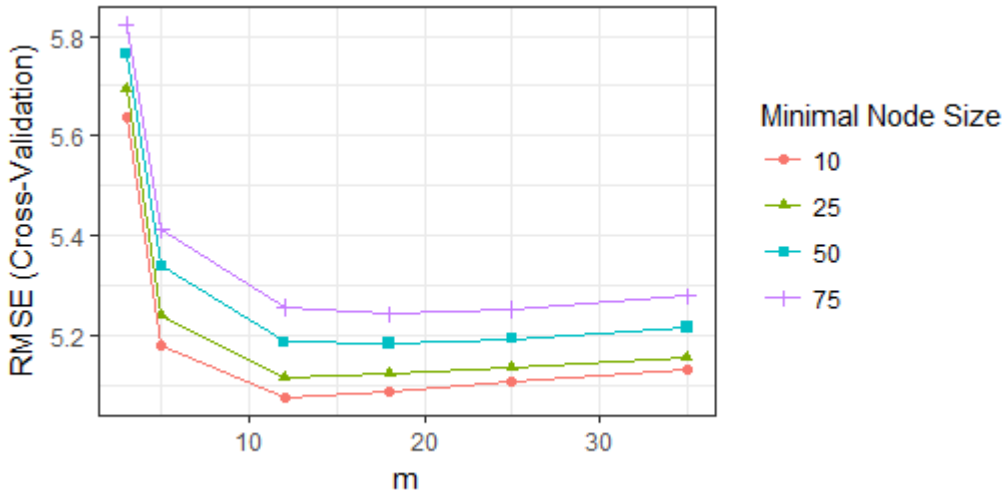


Рисунок 3.35 – График значений $RMSE$ на перекрестной проверке для случайного леса

Здесь в качестве оптимизирующего значения представлена метрика $RMSE = \sqrt{MSE}$. Минимальное значение $RMSE = 5.074186$ достигалось при $m = 12$ и $MN = 10$.

Построенная модель с указанными параметрами на всей обучающей выборке и проверка на тестовой позволило достичь следующих значений метрик:

$$MSE_{tr} = 8.266604, \quad MSE_{ts} = 22.33565$$
$$MAE_{tr} = 0.04369142, \quad MAE_{ts} = 0.1398928 \tag{3.13}$$

Видна высокая степень переобучения модели из-за разницы оценок $MSE_{ts} > MSE_{tr}$ более, чем в 2 раза. Тем не менее, итоговая модель показала неплохой результат в сравнении с линейной регрессией.

Для построения регрессионного градиентного бустинга была выбрана техника последовательного перебора гиперпараметров с помощью перекрестной проверки. Ее суть состоит в том, чтобы последовательно перебирать указанные параметры по блокам:

1. Зафиксировать коэффициент скорости обучения η на определенном уровне и перебирать количество итераций бустинга N. В данном случае, скорость обучения была зафиксирована на более высоком уровне, чем в стандартных условия, а также бралось меньшее количество деревьев. Все иные параметры фиксировались по умолчанию. Это делалось для оптимизации скорости расчетов:

$$\eta = 0.1 \quad N \in \{50, 100, 150, 200, 300, 400, 500\} \tag{3.14}$$

Подбор оптимального количества деревьев производился с помощью десятиблочной перекрестной проверки. График ошибки представлен на рис. 3.36.

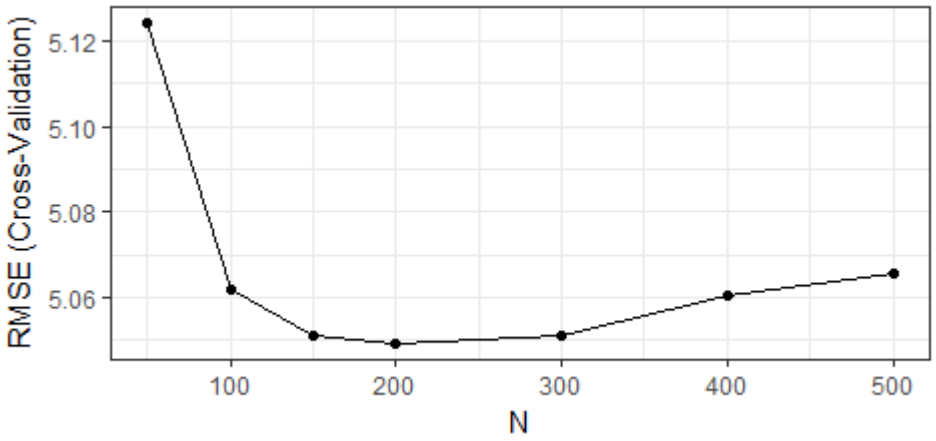


Рисунок 3.36 – График значений $RMSE$ на первом этапе настройки XGBoost (кросс-валидация)

Минимальное значение $RMSE = 5.048955$ было достигнуто при $N = 200$.

2. Фиксируются значения η и N. В рамках расчетов значения $\eta = 0.1$ и $N = 200$. Производился решетчатый поиск для параметров бустинга, которые ограничивают сложность деревьев – это максимальная глубина деревьев MD и минимальное количество объектов в листе дерева MC:

$$MD \in \{3, 4, 6, 8, 10\} \quad MC \in \{1, 10, 25, 50, 100\} \tag{3.15}$$

Также, как и на первом этапе использовалась десятиблочная кросс-валидация и выводилась оценка RMSE, показанная на рис. 3.37.

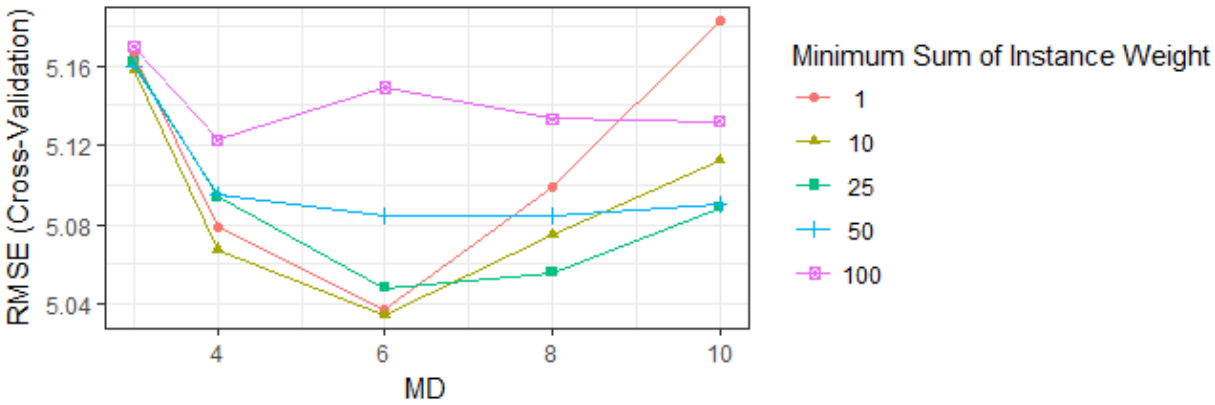


Рисунок 3.37 – График значений $RMSE$ на втором этапе настройки XGBoost (кросс-валидация)

На рисунке 3.37 MC обозначено как «Minimum Sum of Instance Weight». Минимальное значение $RMSE = 5.048955$ достигалось при $MD = 6$ и $MC = 10$.

3. На этом этапе фиксировались значения $\eta = 0.1$, $N = 200$, $MD = 6$ и $MC = 10$. Побиралась заключительная группа параметров, отвечающая за формирование подвыборок. По сути, здесь реализовались идеи Лео Бреймана и создавалась модель стохастического градиентного бустинга – настраивалась доля объектов выборки δ для каждой итерации, а также доля используемых признаков c:

$$\delta \in \{0.3, 0.5, 0.7, 0.9, 1\} \quad c \in \{0.5, 0.7, 0.9, 1\} \tag{3.16}$$

Для десятиблочной перекрестной проверки результат выводится на рис. 3.38.

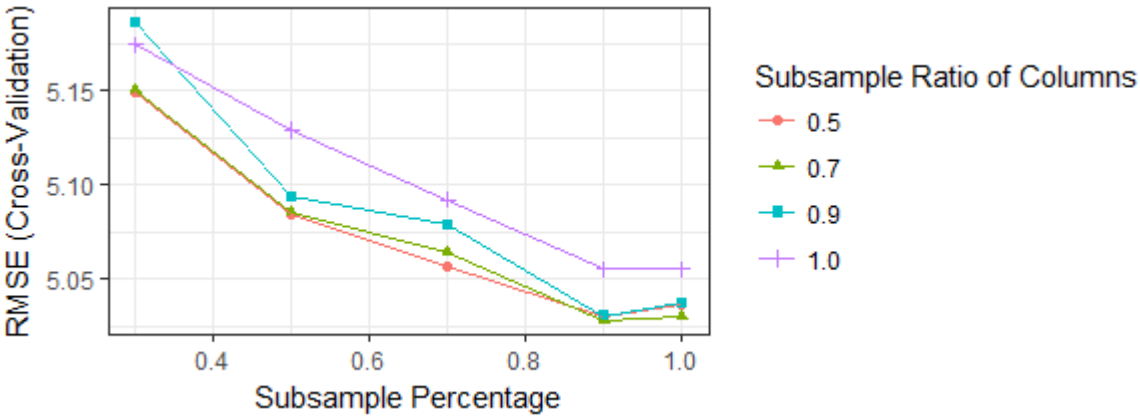


Рисунок 3.38 – График значений $RMSE$ на третьем этапе настройки XGBoost (кросс-валидация)

На рисунке 3.38 δ обозначено как «Subsample Percentage», c – как «Subsample Ratio of Columns». Оптимальное значение $RMSE = 5.027795$ при $\delta = 0.9$ и $c = 0.7$.

4. Подобранная модель была обучена на всех данных и проверялась на тестовом периоде. Отсюда значения метрик:

$$MSE_{tr} = 15.42815, \qquad MSE_{ts} = 22.16758$$
$$MAE_{tr} = 0.0008282764, \qquad MAE_{ts} = -0.130474 \tag{3.17}$$

Видно, что переобучение модели есть, но оно не столь значительно в сравнении с алгоритмом случайного леса. В целом характеристики модели лучше, чем у всех остальных методов.

3.3.3. Экономическая интерпретация важных переменных

Согласно методикам, описанным в разделах 2.3 и 2.4. был сделан ряд предположений о важности нескольких переменных, которые следует подтвердить некоторыми выкладками. Речь идет прежде всего о переменных «Количество товарных позиций, взаимозаменяемых по цене», «Количество продаж в товарной группе», «Количество чеков» и «Температурный режим (среднедневная температура)».

При реализации регрессионного моделирования ненулевого спроса $\widehat{D}$ были также сделаны выводы об указанном выше предположении, которое легло в основу смысла переменной «Количество товарных позиций, взаимозаменяемых по цене» N_{SKU}. Линейная модель $\widehat{D}$ была оценена с двумя вариантами переменной N_{SKU}: рассчитанной для всех товарных позиций и рассчитанной для всех товарных позиций с действующей промо-акцией (снижение цены) на момент прогнозирования по отношению к товару. Коэффициенты полученной модели линейной регрессии равны -0,0077 и -0,0071 (при среднедневном спросе по обучающей выборке 5,96 шт.; медианном спросе – 2 шт.). Данный результат можно интерпретировать следующим образом: при

увеличении количества N_{SKU} на 1 единицу в рамках ценового диапазона $(pr_i,\ pr_i + h]$, в котором находится товар на который прогнозируется спрос, оценка спроса $\widehat{D}$ снижается на 0,0077 единиц (на 0,0071 в случае акционного товарного соседства). В условиях, когда в рамках одного ценового диапазона может продаваться не менее 20-30 товарных позиций это может привести к 2,5% - 4% снижению спроса на товар в среднем. Это позволяет в дальнейшем выводить рекомендации к ассортиментной политике на розничном предприятии.

Указанная в разделе 2.4 методика прогнозирования временных рядов использовалась для моделирования трех ключевых переменных (см. раздел 3.2), которые имеют потенциальный эффект на спрос: иерархически связанный со спросом параметр – количество продаж в товарной группе; параметр, определяющий емкость спроса в рамках розничной точки – количество чеков; параметр, который имеет сильное влияние на товары повседневного спроса – температурный режим. Линейные модели, которые легли в основу прогноза спроса, позволяют оценить влияние указанных показателей на покупательский спрос (таблица 3.12).

Таблица 3.12

Коэффициенты линейных моделей покупательского спроса для ключевых переменных

Полное наименование ключевых переменных	**К-ты логистической модели**	**Связанная переменная (мультипликативно)**	**К-ты линейной модели**
Количество продаж в товарной группе	0,000720	Лаг L1	0,000306
		Лаг L2	0,000099
		Лаг L3	0,000070
		Лаг L4	0,000198
		Лаг L5	0,000160
		Лаг L6	-0,000028
		Лаг L7	-0,000062
Количество чеков	0,000029	-	-0,000105
Температурный режим (среднедневная температура)	-0,001557	-	-0,004793

Согласно таблице 3.12 видно, что с ростом количества продаж в товарной группе растет и вероятность ненулевого спроса, а связь этого показателя с величиной спроса определяется через мультипликативное отношение с лагами спроса. Например, при значении количества продаж равном 1000 логарифм шансов ненулевого спроса увеличивается в 0.72 раза, а сама величина спроса увеличивается на 0,306 от первого лага в штуках (не берем во внимание остальные лаги от значений спроса). Это показывает четкую положительную взаимосвязь между общим масштабом продаж и значениями индивидуального спроса. Интересную интерпретацию имеет взаимосвязь показателя количества чеков и оценки спроса: при росте показателя растет и вероятность покупки, но при этом падает сама величина спроса. То есть логарифм шансов ненулевого спроса увеличивается в 0,000029 раза, а сама величина уменьшается на 0,000105 штуку. Опять же, при количестве чеков равном 3800, что примерно равняется среднему

количеству чеков в магазине, это имеет более значительное изменение. Подобное поведение характерно для данного вида товаров, когда при большом наплыве покупателей спрос стабилизируется, но при этом он замещается другими продуктами, имеющими схожие потребительские характеристики. Рост среднедневной температуры ведет как к снижению вероятности покупки, так и снижению величины спроса: на каждый 1 градус уменьшение логарифма шансов в 0,001557 раза, а величина уменьшается на 0,004793 штуку. Стоит отметить, что с учетом сложности модели, данная интерпретация не является исчерпывающей, но позволяет судить о важности указанных переменных.

3.3.4. Расчет и оценка итогового прогноза спроса

Перед тем как приступить к расчету итогового прогноза спроса комбинировались значения как для классификации, так и для регрессии. Для этого выбиралось простое арифметическое среднее, ввиду того, что это простой в применении и устойчивый метод.
Усреднение для оценки вероятности ненулевого спроса выглядело следующим образом:

$$P_k(y \neq 0) = \sum_{i=1}^{m} p_{k_i} /m = \sum_{i=1}^{3} p_{k_i} /3, \quad (3.18)$$

где $m = 3$ ввиду использования 3-х моделей оценки вероятности, потому как выбирались методы логистической регрессии, случайного леса и градиентного бустинга. Это сделано ввиду того, что значение корреляции между результатами логистической регрессии и логистической регрессии с регуляризацией высоко, а качество регрессии с регуляризацией хуже.

Таблица 3.13

Корреляция между оценками вероятности ненулевого спроса по разным методам

Результаты оценки вероятности	Логистическая регрессия	Логистическая регрессия с регуляризацией	Случайный лес	Градиентный бустинг
Логистическая регрессия	1.0000000	0.9885366	0.9602946	0.9630219
Логистическая регрессия с регуляризацией	0.9885366	1.0000000	0.9513350	0.9485526
Случайный лес	0.9602946	0.9513350	1.0000000	0.9859474
Градиентный бустинг	0.9630219	0.9485526	0.9859474	1.0000000

На рис. 3.39 выводится ROC-кривая по качеству ансамбля на тестовой выборке в сравнении с ROC-кривой по лучшему выявленному методу (градиентный бустинг).

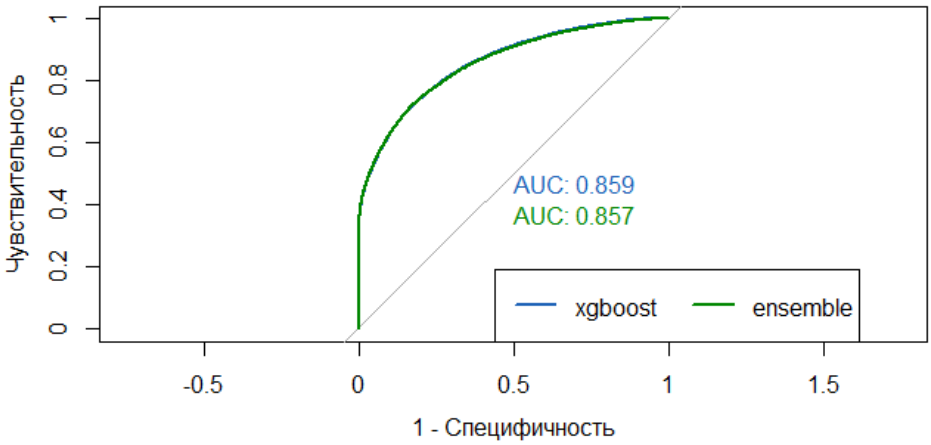


Рисунок 3.39 – ROC-кривые по лучшему методу и комбинации

Видно, что ансамбль не превзошел по качеству лучший метод. При этом, при переборе других комбинаций также не была достигнута оценка $AUC > 0.859$. Поэтому для итоговой оценки, первоначально, выбирался результат по методу градиентного бустинга.

Для регрессионной оценки $\widehat{D}$ также выводится ансамбль прогнозов по принципу:

$$\widehat{D}_k = \sum_{j=1}^{l} d_{k_j} / l = \sum_{j=1}^{3} d_{k_j} / 3 \qquad (3.19)$$

Матрица корреляций регрессионных оценок определяет участие рассматриваемых методов:

Таблица 3.14

Корреляция между оценками $\widehat{D}$ по разным методам

Результаты регрессионного моделирования	**Линейная регрессия**	**Линейная регрессия с регуляризацией**	**Случайный лес**	**Градиентный бустинг**
Линейная регрессия	1.0000000	0.9989419	0.9783826	0.9733069
Линейная регрессия с регуляризацией	0.9989419	1.0000000	0.9799865	0.9753739
Случайный лес	0.9783826	0.9799865	1.0000000	0.9883705
Градиентный бустинг	0.9733069	0.9753739	0.9883705	1.0000000

Количество методов $l = 3$ ввиду того, что, аналогично предыдущей ситуации, прогнозы по линейной регрессии и линейной регрессии с регуляризацией сильно коррелированы. В данном случае в качестве базового линейного метода был выбран метод с регуляризацией, так как он работает точнее. После расчета средней по 3 методам была проведена оценка значений метрик на тестовой выборке:

$$MSE_{ts} = 21.77471$$
$$MAE_{ts} = -0.04042693 \qquad (3.20)$$

Видно, что в сравнении с лучшим методом – градиентным бустингом – (3.17) все метрики показывают результат лучше. То есть для регрессионной задачи комбинация методов оказалась оправдана.

Соответственно, для расчета итогового значения прогноза спроса (2.40) рассчитывалось произведение оценки вероятности ненулевого спроса по бустингу и комбинации регрессионных оценок. Результат итоговой метрики на тестовой выборке:

$$MSE_{ts} = 7.990168$$
$$MAE_{ts} = -0.04111812 \quad (3.21)$$

В данном случае под тестовой выборкой понимается полный объем информации, где $D \geq 0$.

Следует отметить, что в ходе экспериментов, предшествующих созданию методологии прогнозирования спроса, которая используется в итоговой версии диссертационной работы, было выявлено, что при применении классического регрессионного подхода оценки MSE смещены. Например, при получении результатов по ансамблю методов – линейная регрессия с регуляризацией, машина опорных векторов и случайный лес – получена оценка $MSE = 8.634084$. Чтобы иметь представление о распределении оценок метрики MSE в предшествующих компьютерных экспериментах, они приводятся в таблице 3.14.

Таблица 3.15

Распределение MSE для методов в классической постановке регрессионного прогнозирования

Метод прогнозирования	MSE на тестовой выборке
Линейная регрессия	9.077873
Гребневая регрессия (ridge)	9.005262
Лассо регрессия	9.038254
Регрессия на опорных векторах	9.169015
Случайный лес	8.96718
Ансамбль	8.634084

Здесь, для уточнения, прогнозирование производилось на всем объеме данных без учета фактора смещения спроса в 0 значениях (т.е. не производилось расчета оценки вероятности ненулевого спроса и нормировки на это значение). Отсюда, оценка (3.21) является более качественной по сравнению с классическим регрессионным подходом, что подтверждено экспериментально.

Кроме того, в ходе дополнительных экспериментов, соответствующих принципу отраженному на рисунке 2.8, были получены более оптимальные значения MSE и MAE:

$$MSE_{ts} = 7.941281$$
$$MAE_{ts} = -0.0311855 \quad (3.22)$$

Данная оценка была получена при замене метода прогнозирования вероятности ненулевого спроса с градиентного бустинга на алгоритм простой логистической регрессии. Следуя принципу Оккама, этот вариант является более предпочтительным для итогового прогноза ввиду того, что логистическая регрессия обладает более простой спецификацией в

сравнении с градиентным бустингом. Следовательно, согласно формуле (2.38) итоговая математическая модель прогнозирования спроса выглядит следующим образом:

$$\hat{y} = p(X) \cdot \left(d_{k_{lin}}(X) + d_{k_{rf}}(X) + d_{k_{gbm}}(X) \right) / 3 \qquad (3.23)$$

где $p(X)$ – модель вероятности ненулевого спроса на основе логистической регрессии, $d_{k_{lin}}(X)$, $d_{k_{rf}}(X)$ и $d_{k_{gbm}}(X)$ – модели регрессионной оценки спроса на методов основе линейной регрессии с регуляризацией, случайного леса и градиентного бустинга соответственно.

Для дополнительной оценки итогового прогноза спроса проводится визуальная оценка для некоторых товарных единиц (рис. 3.40), а также рассчитывается оценка MSE и MAE.

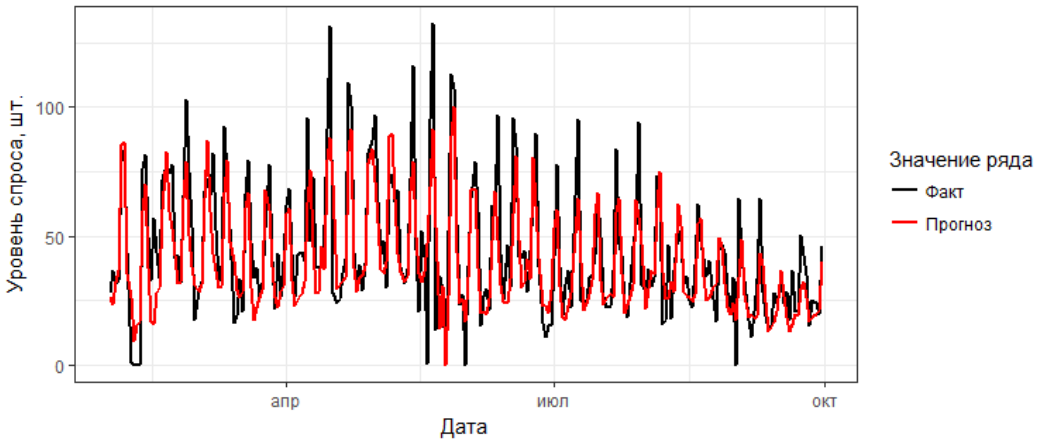


Рисунок 3.40 – График прогноза первого товара

По рисунку 3.40 видно, что в целом, прогноз отлавливает все колебания в спросе, тем не менее существуют участки спроса, в которых идет занижение прогноза. Значение $MSE = 222.4479$, $MAE = 4.426199$, что при таких высоких средних значениях спроса на данный товар является хорошим результатом.

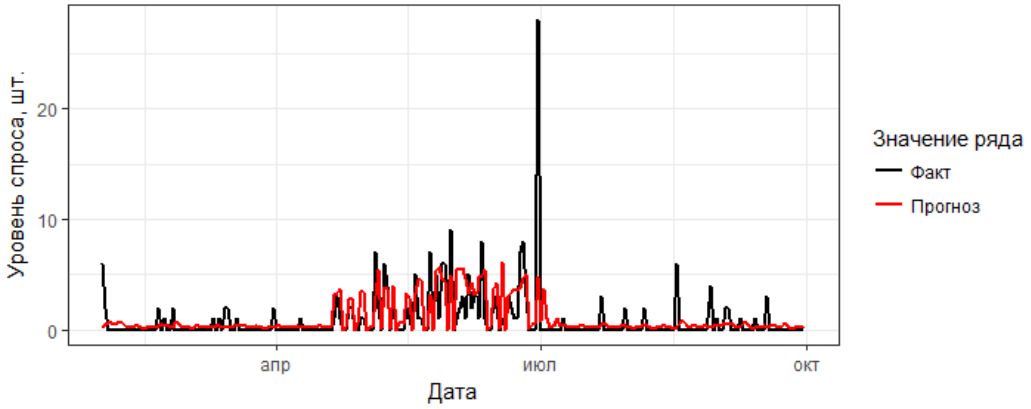


Рисунок 3.41 – График прогноза второго товара

На рисунке 3.41 анализируется товар с нерегулярным спросом. Видно, что система прогнозирования адекватно реагирует на изменения среды и увеличивает прогноз до приемлемого уровня. Есть большой скачок спроса, который является оптовой покупкой, что в

полной мере не учитывается системой прогнозирования. Значение $MSE = 4.230116$ довольно высокое из-за значимого выброса, $MAE = -0.1438258$, что в целом является хорошим результатом.

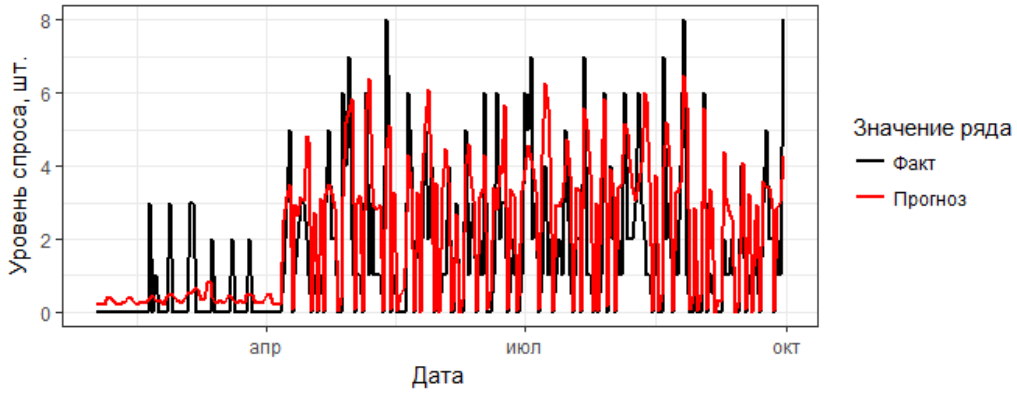


Рисунок 3.42 – График прогноза третьего товара

Рассматривается товар с высокой вариабельностью спроса (рис. 3.42). Несмотря на неслаженный характер целевого значения, прогноз «вылавливает» колебания достаточно корректно. Исключение составляют небольшие значения спроса в начале временного периода. $MSE = 2.470576$ и $MAE = -0.5929248$.

В ходе проведения дополнительной оценки системы прогнозирования на отдельных товарах было выявлено:

- Прогнозные значения в целом попадают в тренд, циклическую и сезонные компоненты фактического ряда. Направление прогноза угадывается довольно точно.
- Есть определенные колебания в оценках, которые, в приложении к формированию реального заказа, сглаживаются с помощью агрегации до периода поставки, а также с помощью страхового запаса. Страховой запас может быть оценен исходя из дисперсии ошибки прогноза.
- Существуют определенные серьезные выбросы, связанные в основном с оптовыми и мелкооптовыми закупками в розничном магазине. Такие случаи должны быть исключены в дальнейшей оценке прогнозной системы.

3.4 Реализация программного комплекса прогнозирования спроса на языке R

Язык R – мультипарадигамальный язык программирования для статистической обработки данных и визуализации. Язык R активно используется для исследований в академической среде и в последнее время получает все большую распространенность для решения бизнес-задач. Особенностью языка является удобство написания высокоуровневых решений на его основе. При этом освоение языка довольно быстро осуществляется и непродвинутым пользователем.

Мощным инструментами для разработки приложений также обладают специальные функциональные расширения для языка R – пакеты, количество которых растет с каждым годом активного использования языка [105]. С начала 2000-х гг. заметна повсеместность использования языка программирования R, предназначенного для анализа данных и статистического моделирования процессов, для решения задач в области прогнозирования. Из-за открытости и гибкости математических методов, применяемых в данном программном продукте, он используется в системах поддержки принятия решений мировых розничных компаний таких как Walmart [116]. Разработка методологии моделирования при этом ведется специалистами самой розничной компании без привлечения третьего лица.

Система прогнозирования – это прежде всего система обработки и анализа данных, на выходе которой исследователь имеет результат в виде модели прогнозирования и предсказания целевой переменной. Подобный класс задач решает такая область методов как Data Mining [60]. В данном контексте разработанная система прогнозирования имеет следующие этапы:

1. Постановка задачи прогнозирования;
2. Сбор данных;
3. Подготовка (предобработка) данных;
4. Выбор подходящего типа модели прогнозирования (линейная регрессия, деревья решений, нейронные сети или т.п.);
5. Подбор метапараметров модели и алгоритма обучения;
6. Обучение модели на данных;
7. Анализ качества модели;
8. Вывод результата (предсказания) в систему поддержки принятия решений на предприятии.

Дополнительное введение последнего пункта связано с тем, что система прогнозирования является частью общей системы управления предприятием.

С точки зрения инструментов разработанного решения, архитектура системы прогнозирования выглядит как на рис. 3.43 [68].

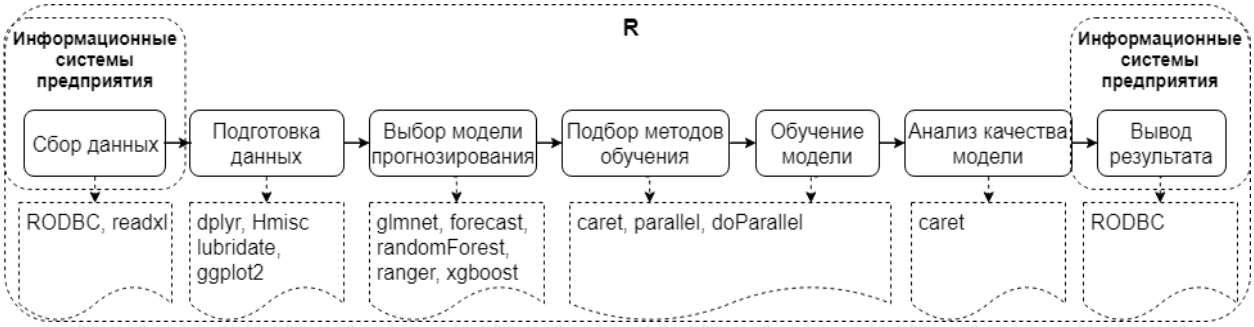


Рисунок 3.43 – Архитектура системы прогнозирования на языке R (с указание доп. пакетов)

Как видно, возможности языка R охватывают все этапы работы системы прогнозирования. При этом, при подготовке решения не обошлось без использования пакетных расширений, широко распространённых среди разработчиков в среде R. Условно их можно разделить на 4 группы:

- Пакеты, которые осуществляют функции интеграции для сбора данных: *RODBC* (для интеграции с SQL-базами данных), *readxl* (для сбора данных с внешних Excel-файлов).
- Пакеты, которые осуществляют функции обработки и подготовки данных к моделированию: *dplyr* (облегчает фильтрацию, агрегацию и иные манипуляции с данными), *lubridate* (облегчает работу с временными данными), *Hmisc* (обработка пропущенных значений), *ggplot2* (позволяет визуализировать необходимые связи в данных).
- Пакеты, которые отвечают за создание и оценку качества моделирования: *glmnet* (линейные модели с регуляризацией), *forecast* (модели временных рядов), *randomForest* (реализация метода «случайный лес»), *ranger* (реализация вероятностного и классического методов «случайный лес»), *e1071* (реализация нескольких методов, в частности машины опорных векторов), *xgboost* (реализация градиентного бустинга), *caret* (отвечает за организацию моделирования и оценку качества результата);
- Пакеты, которые отвечают за повышение эффективности реализуемого программного комплекса: *parallel, doParallel* (позволяет ускорить процедуры за счет параллелизации процессов).

Использование пакетов для языка R позволило сделать решение всеобъемлющим и гибким с точки зрения настройки возможных результатов.

Особенностью данного решения является использование не одной модели прогнозирования, а композиции (ансамбля) нескольких методов. При этом, сам выбор методов подтверждается выводами статистического и визуального анализа, который совершается средствами R. Также проводится процедура разделения выборки на обучающую (*train*) и тестовую (*test*), процедура перекрестной проверки (или кросс-валидации) для подбора метапараметров моделей и само их обучение. Все это позволяет совершить в эффективной высокоуровневой форме пакет *caret*.

Инструментально данный подход также реализуется с помощью R: для ускорения процедуры создания нескольких моделей используются пакеты для параллелизации вычислений (*parallel, doParallel*), а для создания ансамбля – пакет *caret* и его расширение *caretEnsemble*, в котором реализованы функции предназначенные для создания и оценки требуемых композиций.

Полученная система прогнозирования спроса функционально интегрируется с глобальной информационной системой предприятия. По сути модуль на R – это некий микросервис [111], который интегрируется с информационной системой предприятия. В онлайн-режиме и статических отчетах результат прогноза используется для заказа товара у поставщика, планирования показателей продаж (количественных и стоимостных), проработки маркетинговых стратегий и многих других стратегических мероприятий.

Согласно методологии прогнозирования, практическая реализация на языке R по блокам прогнозирования отражена в приложениях следующим образом:

- Прогнозирование временных показателей, влияющих на спрос, на примере температурного режима в Приложении В.
- Моделирование прогноза спроса и комбинация этих прогнозов в Приложении Г.

Разработанное ПО используется в бизнес-процессах сети ООО «Гастроном», а также может быть использовано в розничных сетях схожего формата во всех регионах России. Это делает продукт исследования конкурентоспособным в рамках практической плоскости применения результатов диссертационного исследования.

3.5. Оценка изменений в системе управления товарными запасами

Созданная система прогнозирования спроса не может быть оценена исчерпывающе только с помощью метрик качеств самого прогноза. Так как модель прогнозирования покупательского спроса – это ключевая составляющая автоматизации процессов розничного магазина, в частности процессов товародвижения, то необходимо оценить внедряемую систему с точки зрения экономики торгового предприятия. Для этого была выбрана задача формирования заказа на товар, то есть задача управления товарными запасами [69]. Здесь система прогнозирования выполняет роль системы поддержки принятия решений. Для оценки эффективности внедрения системы прогнозирования для работы автоматического заказа использовались следующие показатели:

- Среднее значение товарных запасов за период:

$$\bar{I} = \frac{\frac{1}{2} \times I_1 + I_2 + \cdots + I_{n-1} + \frac{1}{2} \times I_n}{n-1}, \tag{3.24}$$

где $\bar{I}$ – средние товарные запасы за период n, которые могут быть выражены как в рублях так и в штуках.

- Уровень товарных запасов в днях – количество дней, в течение которого будет израсходован текущий товарный запас с учетом средних продаж за предыдущий период:

$$L_i = \frac{I_i}{\overline{D}_i}, \tag{3.25}$$

где L_i – уровень -го запаса в днях (обеспеченность), I_i – остаток запаса для i-го товара в анализируемый момент времени, $\overline{D}_i$ – средний товарооборот или количество проданного товара в зависимости от природы показателя. $\overline{D}_i$ вычисляется как D_i/t, где t – это количество периодов для расчета средних продаж.

– Показатель товарооборачиваемости – время обращения среднего запаса за определенный период:

$$Kq_i = \frac{D_i}{\overline{I}_i} \tag{3.26}$$

Был проведен расширенный анализ изменений в характеристиках системы автозаказа после замены модуля прогнозирования. Это означает, что все характеристики действующей системы заказа и управления запасами оставались неизменными кроме оценки спроса. Для примера был взят товар с продажами и прогнозом как на рисунке 3.42 и проведены соответствующие расчеты на тестовой выборке.

По сути размер требования определяется как

$$Q = k_s \times \sum_{i=1}^{c} \hat{y}_i, \tag{3.27}$$
$$Q = ceiling(Q)$$

где $\hat{y}_i$ – прогноз спроса на i -й день, $c = 7$ дней, так как действует недельный цикл поставки, k_s – коэффициент страхового запаса, который по умолчанию принимается за $k_s = 1.2$. В заключение выполняется процедура до ближайшего целого - $ceiling$. Для простоты предполагалось, что после формирования заказа поставка осуществляется единовременно.

Динамика уровня товарных запасов (на начало дня) представлена на рисунке 3.44.

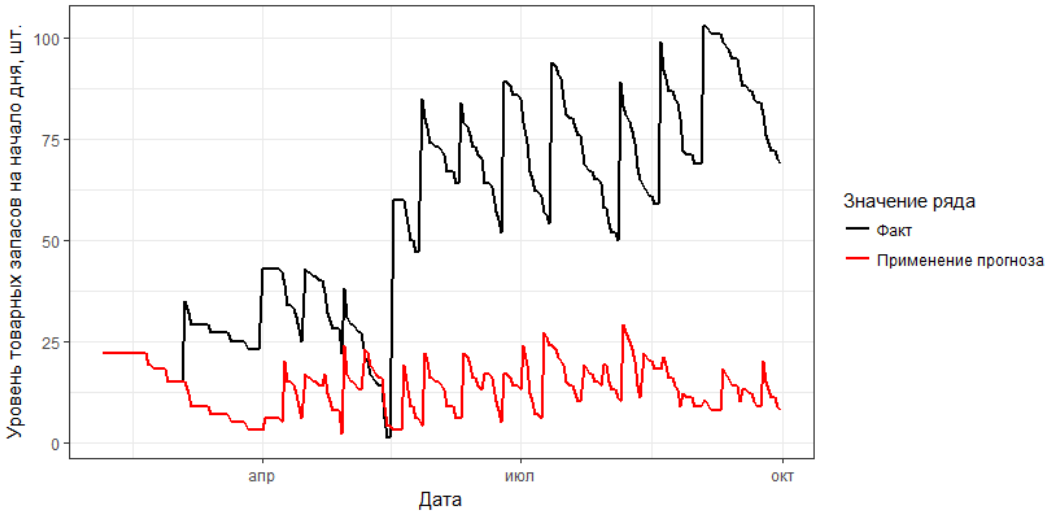


Рисунок 3.44 – Динамика товарных запасов на начало дня

На рисунке вполне наглядно отражен процесс оптимизации товарных запасов для указанного товара. Следует привести характеристики этого процесса на таблице 3.16.

Таблица 3.16

Показатели уровня товарных запасов (количество)

Показатель	Фактическое значение	Значение при применении прогноза	Прирост
Среднее значение товарных запасов, в шт.	53,99	13,57	-74,87%
Уровень товарных запасов в днях (среднее значение), в дн.	44,83	10,07	-77,54%
Показатель товарооборачиваемости (среднее значение), количество раз	0,03	0,13	290,56%

Характеристики сформированы на количественных измерениях: количество продаж и количество остатков. Видно, что внедрение системы прогнозирования при еженедельной поставке серьезно оптимизировало уровень товарных запасов. При этом, что видно по рисунку 3.44, не было факта out-of-stock: весь спрос был удовлетворен.

Следует также оценить изменения в стоимостном эквиваленте (Таблица 3.17).

Таблица 3.17

Показатели уровня товарных запасов (стоимость)

Показатель	Фактическое значение	Значение при применении прогноза	Прирост
Среднее значение товарных запасов, в руб.	4114,04	1034,03	-74,87%
Уровень товарных запасов в днях (среднее значение), в дн.	31,09	6,98	-77,55%
Показатель товарооборачиваемости (среднее значение), количество раз	0,05	0,19	289,80%

В среднем значение товарных запасов в стоимостном эквиваленте снизилась на 3 080 рублей ежедневно. Это говорит о серьезной экономии средств на обеспечение запасов, а также о возможности розничному предприятию вложить освободившиеся денежные средства либо в обеспечение тех запасов, которые действительно необходимы, либо на любые другие управленческие решения стратегического и тактического характера.

Исходя из расчетов видно, что внедренная система прогнозирования решает задачу оптимизации товарных запасов. В целом, при внедрении системы прогнозирования в такой связке с автоматическим заказом, удалось в среднем сократить товарные запасы на 9,6%. При этом, на некоторых товарных позициях со снижением эффекта out-of-stock, произошло увеличение продаж. Исходя из этого общие продажи по группе увеличились на 4,1%.

ЗАКЛЮЧЕНИЕ

В диссертационной работе построены экономико-математические модели прогнозирования ключевых показателей розничной торговли и потребительского спроса, которые снижают издержки предприятия:

1. Разработана методология построения математических моделей прогнозирования ключевых показателей, связанных с основными бизнес-процессами розничной торговли, основанная на инструментах анализа временных рядов. В контексте самостоятельного использования данная методология позволяет прогнозировать большое количество тренд-сезонных процессов в розничной торговле, а также может быть использована при разработке планов для показателей.
2. Разработана модель прогнозирования покупательского спроса на основе методов машинного обучения и показателей, характеризующих поведение покупателя. Предложенная модель используется в основе автоматизированного заказа товара, тем самым минимизируя издержки на обеспечение товарного запаса на предприятии.
3. Разработан программный комплекс на основе языка R, который в виде сервиса работает совместно с системами поддержки принятия решений на предприятиях сферы розничной торговли. В ходе разработки сервиса использовались расширенные возможности R, основанные на подключении дополнительных пакетов.
4. Разработанный комплекс математических моделей в виде сервиса (набора сервисов), реализованных на языке R, готов к внедрению на любых предприятиях розничной торговли, которые имеют достаточную историю продаж.
5. Применение разработанного сервиса в рамках компании ООО «Гастроном» позволило снизить средние товарные запасы на 9,6%, при этом увеличив продажи товара на 4,3%.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Абрютина, М. С. Экономический анализ товарного рынка и торговой деятельности: учебник / М. С. Абрютина. – М.: Дело и сервис, 2010. – 462 с.
2. Айвазян, С. А. Методы эконометрики: учебник / С.А. Айвазян. – М.: Магистр: ИНФРА-М, 2010. – 512 с.
3. Айвазян С.А., Фантаццини Д. Эконометрика-2. Продвинутый курс с приложениями в финансах. – М.: Магистр: Инфра-М, 2014. — 944 с.
4. Акобир, Ш. Деревья решений – общие принципы работы [Электронный ресурс] BaseGroup – Labs, 2017 – Режим доступа: https://basegroup.ru/community/articles/description
5. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование: Учебник. — М.: Финансы и статистика, 2001. — 228 с.
6. Баль А.В., Логиновский О.В. Автоматизированный заказ высокооборачиваемых товаров с низкими сроками годности с использованием почасовых продаж// Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2015. – Выпуск № 1. Том 15. – С. 21-25
7. Баринова О. В., Вальков А. С., Воронцов К. В., Громов С. А., Ефимов А. Н., Чехович Ю. В. Система прогнозирования потребительского спроса Goods4Cast. [Электронный источник] – Режим доступа: http://www.ccas.ru/frc/papers/voron05goods4cast.pdf
8. Бегутова, С. В. Использование методов интеллектуального анализа данных для оценки риска неуплаты таможенных платежей // Вестник ОГУ. – 2010. – №1 (107). – С.98-102
9. Безбородова, Т.М., Дюжева, М.Б. Управление предприятиями торговли. Учебное пособие – Омск: РГТЭУ, 2013. – 340 с.
10. Берман, Барри, Эванс, Джоэл Р. Розничная торговля: стратегический подход, 8-е издание. – М.: Издательский дом Вильямс, 2003. – 1184 с.
11. Большой экономический словарь / А.Н. Азрилиян. – М.: Институт новой экономики, 2002. - 469 с.
12. Борзых Д. А., Демешев Б. Б. Эконометрика в задачах и упражнениях. – М. : УРСС, 2017. – 304 с.
13. Валевич Р.П., Давыдова Г.А. Экономика торговой организации. Учебное пособие. – Минск: Вышэйш. шк., 2008. – 371 с.
14. Вестник Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE). Вып. 6 [Электронный ресурс]: Сб. науч. статей / Отв. ред.: П. М. Козырева. - М.: НИУ ВШЭ, 2016. – Режим доступа: https://www.hse.ru/data/2016/07/28/1118935935/Vestnik%20RLMS-HSE_2016.pdf

15. Ветров Д.П., Кропотов Д.А. Байесовские методы машинного обучения, учебное пособие по спецкурсу, 2007 [Электронный ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D0%BC%D0%BC%D0%BE
16. Виноградова С.Н. Коммерческая деятельность. Учебник. — 2-е изд., испр. — Минск: Выш. шк., 2012. — 288 с.
17. Воронцов, К.В.. Лекции по методу опорных векторов [Электронный ресурс] – Режим доступа: http://www.ccas.ru/voron/download/SVM.pdf (41)
18. Воронцов, К.В. Лекции по алгоритмическим композициям [Электронный ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf
19. Голдратт Э., Эшколи А., Лир Дж. Б. Я так и знал! Розничная торговля и Теория ограничений— М.: Альпина Паблишер, 2018. — 168 с.
20. Грицай, А. А. Интеллектуальная система управления запасами forecast now! // Инновации. – 2014. №2 (184). [Электронный ресурс] – Режим доступа: http://cyberleninka.ru/article/n/intellektualnaya-sistema-upravleniya-zapasami-forecast-now (дата обращения: 15.01.2016)
21. Груздев, А. В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес. – М.: ДМК Пресс, 2018. – 642 с.
22. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение : пер. с анг. А. А. Слинкина. – М.: ДМК Пресс, 2017. – 652 с.
23. Демешев Б. Б., Тихонова А. С. Прогнозирование банкротства российских компаний: межотраслевое сравнение // Экономический журнал ВШЭ. – 2014. – №3. – С.359-386
24. Денисов Д.В., Смирнова Д.К. Применение метода случайных лесов для оценки резерва произошедших, но еще не заявленных убытков страховой компании // International Journal of Open Information Technologies. – 2016. – №7. – С.45-51
25. Денисов Н. В. Вербальная модель формирования и развития потребительского спроса // Социально-экономические явления и процессы. – Тамбов. – 2011. – № 5-6. – С.83-86
26. Денисов Н.В., Золотухин Д.Н. Детерминанты «прогрессивного» потребительского спроса // Социально-экономические явления и процессы. – Тамбов. – 2013. – № 3. – С.54-59
27. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. / пер. с англ. Мастицкий С.Э. – М.: ДМК-Пресс, 2016 г. – 460 с.
28. Домингос, П. Верховный алгоритм. Как машинное обучение изменит наш мир. – М.: Манн, Иванов и Фербер, 2016. — 336 с.
29. Доугерти, К. Введение в эконометрику: Пер. с англ. — М.: ИНФРА-М, 1999. — 402 с.
30. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 1. В 2-х кн. М.: Финансы и статистика, 1986. 366 с.

31. Дуброва, Т. А. Статистические методы прогнозирования: Учеб. пособие дом вузов. - М.: ЮНИТИ-ДАНА, 2003. – 206 с.
32. Жуликов, С. Е. Математическое моделирование краткосрочного прогноза погоды // Вестник Тамбовского университета. Серия Естественные и технические науки. – 2009. – №5-2. – С.1021-1026.
33. Журавлёв, Ю. И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлёв, В.В. Рязанов, О.В. Сенько – М.: Фазис, 2006. – 147 с.
34. Иванов, Г.Г. Экономика торгового предприятия : учебник / Г.Г. Иванов. — М. : Издательский центр «Академия», 2010. — 320 с. ISBN 978-5-7695-5744-6
35. Канторович, Г.Г. Лекции: Анализ временных рядов // Экономический журнал ВШЭ. 2002. №1. – [Электронный ресурс] – Режим доступа: http://cyberleninka.ru/article/n/lektsii-analiz-vremennyh-ryadov-4 (дата обращения: 31.05.2017).
36. Катаева, Н.Н. Характеристика и оценка эффективности мерчандайзинга продуктового магазина // Nauka-rastudent.ru. 2014. № 12. – [Электронный ресурс] – Режим доступа: http://naukarastudent.ru/12/2242 (дата обращения: 18.06.2017)
37. Китова Ольга Викторовна, Колмаков Игорь Борисович, Пеньков Илья Андреевич Метод машин опорных векторов для прогнозирования показателей инвестиций // . 2016. №4. С.27-30
38. Ковалев, К. Логистика в розничной торговле: как построить эффективную сеть/ К. Ковалев, С. Уваров, П. Щеглов. – СПб: Питер, 2007 г. - 272 с.
39. Козерод, Л.А. Экономика торгового предприятия: учебное пособие. – Хабаровск: ДВГУПС, 2012.- 175 с.
40. Котлер Ф., Келлер К. Л. Маркетинг менеджмент. 12-е изд. — СПб.: Питер, 2007. — 816 с.
41. Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. 2-е издание / пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 302 с.
42. Кремер, Н.Ш. Математика для экономистов: от Арифметики до Эконометрики : учеб.-справ. пособие для бакалавров / под ред. Н.Ш. Кремера. – 3-е изд., перераб. и доп. – М. : ИД Юрайт, 2012. – 685 с.
43. Кук, Д. Машинное обучение с использованием библиотеки H2O : пер. с англ. А. Б. Огурцова. – М.: ДМК Пресс, 2018. – 250 с.
44. Куприенко, Н. В. Статистические методы изучения связей. Корреляционно-регрессионный анализ/ под ред. Н. В. Куприенко, О. А. Пономарева, Д. В. Тихонов. СПб. : Изд-во политехн. ун-та, 2008. – 118 с.

45. Кэмерон, Э. К. Микроэконометрика: методы их применения. Книга 1./ Э. К. Кэмерон, П. К. Триведи ; пер. с англ. [Сурен Авакян и др.] ; под науч. ред. Б. Демешева. – М.: Дело, 2015. – 552 с.
46. Леви М., Вейтц Б. А. Основы розничной торговли / Пер. с англ. под ред. Ю. Н. Каптуревского. — СПб: Питер, 1999. — 448 с.
47. Лопатников, Л. И. Экономико-математический словарь: Словарь современной экономической науки. – 5-е изд., перераб. и доп. – М.: Дело, 2003. – 520 с.
48. Лукинский, В. В. Актуальные проблемы формирования теории управления запасами : монография / В. В. Лукинский. – СПб. : СПбГИЭУ, 2008. – 213 с.
49. Лысенко, Ю. В. Экономика предприятия торговли и общественного питания : [учеб. пособие для бакалавров и специалистов] : [гриф УМО] / Ю. В. Лысенко, М. В. Лысенко, Э. Х. Таипова. — СПб.: Питер, 2013. – 364 с.
50. Магнус Я.Р., Катыше П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. — 6-изд., перераб. доп. - М.: Дело, 2004. - 576 с.
51. Маккинли, У. Python и анализ данных / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 482 с.
52. Мерков, А. Б. Распознавание образов: Введение в методы статистического обучения. – М.: Едиториал УРСС, 2011. – 256 с.
53. Мишулина, О. А. Статистический анализ и обработка временных рядов : учеб. пособие для студентов вузов / О. А. Мишулина ; М-во образования Рос. Федерации, М-во Рос. Федерации по атом. энергии, Моск. инж.-физ. ин-т (гос. ун-т), Экон.-аналит. ин-т, Каф. экон. динамики. - М. : Моск. инж.-физ. ин-т (гос. ун-т), 2004. - 178 с.
54. Мюллер А., Гидо С. Введение в машинное обучение с помощью Python. – М.: O'Reilly Media, 2017. — 392 с.
55. Насибуллина, З.З. О применении нейронных сетей в экономике и перспективы их развития // Материалы VIII Международной студенческой электронной научной конференции «Студенческий научный форум». – [Электронный ресурс] – Режим доступа: www.scienceforum.ru/2017/2484/32073 (дата обращения: 12.06.2017)
56. Никитин, А. П. Анализ транзакционных данных и определение количественных критериев лояльности клиентов // Экономика. Налоги. Право. –2012. – №2. – С.113-124.
57. Об алгебраических методах в задачах распознавания и классификации. Распознавание, классификация, прогноз / Ю.И. Журавлев — М.: Наука, 1988. — Т. 1. — С. 9-16.
58. Орлов, А.И. Эконометрика : учебник для вузов / А.И. Орлов. — Ростов н/Д : Феникс, 2009. — 277 с.

59. Осовский, С. Нейронная сеть для обработки информации / Пер. с польского И.Д. Рудинского. – М.: Финансы и статистика, 2016. – 448 с.
60. Паклин, Н. Б., Орешков, В. И. Бизнес-аналитика: от данных к знаниям (+CO): Учебное пособие. 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.
61. Памбухчиянц, О. В. Технология розничной торговли: Учебник / О. В. Памбухчиянц. - 9-е изд., перераб. и доп. - М. : ИДК Дашков И КО, 2012. - 288 с.
62. Пивкин, К. С. Корреляционный анализ факторов влияния на покупательский спрос розничного магазина как этап формирования модели прогнозирования и управления запасами // Вестник Удмуртского университета. Серия Экономика и право. – 2016. – №3. – С.40-50.
63. Пивкин, К. С. Алгоритм построения линейной модели на панельных данных как этап эконометрического прогнозирования товарного спроса // Вестник Удмуртского университета. Серия Экономика и право. – 2017. – №2. – С.50-59.
64. Пивкин, К. С. Прогнозирование ключевых показателей розничной сети во времени // Вестник Пермского университета. Серия "Экономика" = Perm University Herald. ECONOMY. 2017. – Том 12. – №4. – С.592-608.
65. Пивкин, К. С. Использование математических методов прогнозирования в системе управления товарными запасами / К. С. Пивкин // Математические методы и интеллектуальные системы в экономике и образовании: сб. материалов всероссийской заочной науч.-практ. конф. – Ижевск, 2015. – С. 21-24.
66. Пивкин, К. С. Система управления товарными запасами на предприятии розничной торговли как объект экономико-математического исследования / К. С. Пивкин // Математические методы и интеллектуальные системы в экономике и образовании: сб. материалов всероссийской заочной науч.-практ. конф. – Ижевск, 2016. – С. 67-70.
67. Пивкин, К. С. Постановка задачи прогнозирования спроса как оценки математического ожидания / К. С. Пивкин // Математические методы и интеллектуальные системы в экономике и образовании: сб. материалов всероссийской заочной науч.-практ. конф. – Ижевск, 2017.
68. Пивкин, К. С. Создание системы прогнозирования с помощью языка R на примере розничного предприятия / К. С. Пивкин // Перспективные информационные технологии (ПИТ 2017): труды Международной научно-технической конференции / под ред. С.А. Прохорова. – Самара: Издательство Самарского научного центра РАН, 2017. – С. 381-385.
69. Пивкин, К. С. Прогноз покупательского спроса как элемент системы управления товарными запасами / К. С. Пивкин // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2017» / Отв. ред. И.А. Алешковский, А.В. Андриянов, Е.А. Антипов.

[Электронный ресурс] — М.: МАКС Пресс, 2017. – Режим доступа: https://lomonosov-msu.ru/archive/Lomonosov_2017/data/section_13_10999.htm

70. Пивкин, К.С. Кластеризация временных рядов как этап прогнозирования покупательского спроса / Математическое и компьютерное моделирование в экономике, страховании и управлении рисками: материалы VI Международной молодёжной научно-практической конференции. – Саратов: ООО Изд-во «Научная книга», 2017. –264 с. – С. 165-169
71. Писарева, О. М. Методы социально-экономического прогнозирования: Учебник ГУУ – НФПК. – М.: Высшая школа, 2003. – 395 с.
72. Рашка, С. Python и машинное обучение / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.
73. Розанова, Н. М. Макроэкономика: учебник для магистров / Н. М. Розанова. М.: Изд-во Юрайт, 2013. – 813 с.
74. Садовникова Н.А., Шмойлова Р.А. Анализ временных рядов и прогнозирование. Вып. 3: Учебно-методический комлекс. – М.: Изд. центр ЕАОИ, 2009. — 264 с.
75. Светуньков, И. С. Методы социально-экономического прогнозирования. В 2 т. Т. 2. Модели и методы : учебник и практикум для академического бакалавриата / И. С. Светуньков, С. Г. Светуньков. — М. : Издательство Юрайт, 2016. — 447 с.
76. Светуньков, И. С. Методы и модели социально-экономического прогнозирования : учебник и практикум для академического бакалавриата. В 2-х т. Т. 1. Теория и методология прогнозирования / И. С. Светуньков, С. Г. Светуньков. — М. : Издательство Юрайт, 2014. — 351 с.
77. Семенова Ю.А., Батукова Л.Р. Характеристика параметрических моделей оценки риска банкротства //Актуальные проблемы авиации и космонавтики – Красноярск: ФГБОУ ВО СГУНиТ. – 2010. – №6. – С.127-128
78. Семёнычев В. К., Семёнычев Е. В. Параметрическая идентификация рядов динамики: структуры, модели, эволюция: монография. – Самара: Изд-во «СамНЦ РАН», 2011. – 364 с.
79. Ситуация и тенденции: российский рынок алкоголя // Исследование аналитической группы Nielsen. 2015. [Электронный ресурс] – Режим доступа: http://www.nielsen.com/ru/ru/insights/news/2015/Alcohol-market-trends-2015-Russia.html
80. Смоглюков, Н.И. Математические методы прогнозирования: Учебно-метод. пособие / Н.И. Смоглюков. – Мн.: БГЭУ, 2005. – 84 с.
81. Соловьева Ю.С, Грекова Т.И. Моделирование экономических процессов с применением нейросетевых технологий // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2009. – №1 (6). –С.49-58

82. Сухарев, М. Г. Методы прогнозирования. Учебное пособие — М.: РГУ нефти и газа, 2009 г. – 208 с.
83. Тененев, В.А., Паклин, Н.Б. Гибридный генетический алгоритм с дополнительным обучением лидера // Интеллектуальные системы в производстве. - 2003. - № 2. - Ижевск: Изд-во ИжГТУ, 2003. - С. 181-206.
84. Торговое дело: экономика и организация: Учебник/Под общ. ред. проф. Л.А. Браги Трохинова А. А. Анализ эффективности деятельности предприятия ресторанно-гостиничного бизнеса [Текст] / А. А. Трохинова, Т. А. Карапетян // Экономическая наука сегодня: теория и практика : материалы V Междунар. науч.–практ. конф. (Чебоксары, 3 дек. 2016 г.) / редкол.: О. Н. Широков [и др.]. Чебоксары: ЦНС «Интерактив плюс», 2016. – С. 95–101.
85. Торговое дело: экономика и организация: Учебник/Под общ. ред. проф. Л.А. Брагина и проф. Т.П. Данько. - М.: ИНФРА-М, 1997. - 256 с.
86. Уоллас Т., Сталь Р. Планирование продаж и операций. Практическое руководство. СПб.: Питер, 2010. - 272 с.
87. Учебник СтатСофт по статистике. Раздел: Анализ временных рядов [Электронный ресурс] – Режим доступа: http://statsoft.ru/home/textbook/default.htm.
88. Фёрстер Э., Ренц Б. Методы корреляционного и регрессионного анализа: руководство для экономистов. М.: «Финансы и статистика», 1983. – 304 с.
89. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.
90. Хайкин, С. Нейронные сети: полный курс, 2е издание. : пер. с англ. – М.: ИД Вильямс, 2006. – 1104 с.
91. Цыплаков, А. Введение в моделирование в пространстве состояний // Квантиль. 2011. – № 9, – С. 1-34.
92. Четвериков, А.А. Линейные модели со смешанными эффектами в когнитивных исследованиях // Российский журнал когнитивной науки. – 2015. Т. 2, –№ 1. – С. 41–51.
93. Чистяков, С. П. Случайные леса: обзор // Труды Карельского научного центра РАН –2013. – №1 – С. 117–136
94. Чучуева, И.А. Модель прогнозирования временных рядов по выборке максимального подобия : автореферат дис. ... кандидата технических наук : 05.13.18 / Чучуева Ирина Александровна; [Место защиты: Моск. гос. техн. ун-т им. Н.Э. Баумана]. - Москва, 2012. - 16 с.

95. Экономика предприятия (торговли и общественного питания): Учебник / С.Е. Метелев, Н.М. Калинина, С.Е. Елкин, В.П. Чижик. – Омск: Издатель Омский институт (филиал) РГТЭУ, 2011. – 474 с.
96. Энциклопедия социологии. Antinazi. 2009 //Интернет-портал "Словари и энциклопедии на Академике [Электронный ресурс] – Режим доступа: http://dic.academic.ru/dic.nsf/socio/1516
97. Энциклопедия эпистемологии и философии науки./И.Т.Касавин – М.: РООИ Реабилитация. И.Т., 2009. – 1248 с.
98. Ясницкий Л.Н. Интеллектуальные системы. – М.: Лаборатория знаний, 2016. – 221 с.
99. Bartmann D., Bach M. F. Inventory Control: Models and Methods. Publisher: Springer Science & Business Media, 1992. – 252 p.
100. Breiman, L. Statistical Modeling: The Two Cultures. Statistical Science? 2001, Vol. 16, N. 3, P. 199–231
101. Caro F., Gallien J. Inventory Management of a Fast-Fashion Retail Network. Article in Operations Research 58(2) · January 2008
102. Cichosz, P. Data Mining Algorithms: Explained Using R. Published: John Wiley & Sons, 2015. – 716 p.
103. Crum C., Palmatier G. Demand Management Best Practices: Process, Principles and Collaboration (Integrated business management series). Publisher: J Ross Publishing (1 July 2003) – 240 p.
104. Cutler A., Breiman L. RAFT: Random Forest Tool. URL: http://www.stat.berkeley.edu/users/breiman/RandomForests/ (дата обращения 03.06.2017).
105. De Vries, Andrie. On the growth of CRAN packages. [Электронный ресурс] / A. de Vries. – Электрон. текстовые дан. – USA, 2016. – Режим доступа: https://www.r-bloggers.com/on-the-growth-of-cran-packages/, свободный.
106. Domingos, P. A few useful things to know about machine learning. Communications of the ACM. Vol. 55. № 10. – 78-87 P.
107. Durbin, J. and Koopman, S. J. Time Series Analysis by State Space Methods. Oxford: Oxford University Press, 2001. 273 p.
108. Gardner, E. S. Exponential smoothing: the state of the art — part II // International Journal of Forecasting. 2006. Vol. 22(4). P. 637–666.
109. George Athanasopoulos, Rob J Hyndman, Haiyan Song, Doris Wu. The tourism forecasting competition. – International Journal of Forecasting 27(3), 2011
110. Hayfield T., Racine J. S. Nonparametric econometrics: The np package. Journal of statistical software 27 (5), 1-32 P.

111. Kiser, M. Deploying R Models in Production. URL: https://blog.algorithmia.com/deploying-r-models-production-web-services/
112. Kuhn M., Johnson K. Applied Predictive Modeling. Springer, 2013. 600 p. — 203 illus., 153 illus. in color.
113. Li Q., Racine J. S. Nonparametric Econometrics: Theory and Practice. Princeton University Press, 2007 – 746 p.
114. Lichtenstein, N. The Retail Revolution: How Wal-Mart Created a Brave New World of Business. Publisher: Picador; 1 edition (June 8, 2010) – 432 p.
115. Ng, A. Machine Learning Yearning. URL: http://www.mlyearning.org/ (96)
116. How Big Data Analysis helped increase Walmarts Sales turnover? URL: https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109
117. Rob J. Hyndman, Yeasmin Khandakar. Automatic Time Series Forecasting: The forecast Package for R. – Journal of Statistical Software. July 2008, Volume 27, Issue 3.
118. Rumelhart D. E., Hinton G. E., Williams R. J. Learning internal representations by error-propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1. 1986. – 318-362 P.
119. Scott S. L. and Varian H. R., Predicting the Present with Bayesian Structural Time Series. 2013. URL: https://ssrn.com/abstract=2304426 (дата обращения: 05.08.2017).
120. Svetunkov I., Kourentzes N. (February 2015). Complex exponential smoothing. Working Paper of Department of Management Science, Lancaster University 2015:1, 1-31.
121. Svetunkov, I. Complex Exponential Smoothing. A thesis submitted for the degree of Doctor of Philosophy, Lancaster University. 2016.
122. Taylor S. J. and Letham B. Forecasting at Scale. URL: https://facebookincubator.github.io/prophet/static/prophet_paper_20170113.pdf (34)
123. Tayur S., Ganeshan R., Magazine M. Quantitative Models for Supply Chain Management. Publisher: Springer US, 1999. – 885 p.
124. Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28(2): 3-28.
125. Zou H., Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Vol. 67. №2. – 301-320 P.

ПРИЛОЖЕНИЕ А ПЕРЕЧЕНЬ ПЕРЕМЕННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

Общие названия переменных	Уточненное название переменной	Тип переменной	Логистическая регрессия	Логистическая регрессия с регуляр-й	Случайный лес	Градиентный бустинг	Примечание
Код товара	Код товара	Категориальная	-*	-	-	-	Используется для анализа
Лаговая переменная спроса	Лаг L1	Количественная	+	+	+	+	Лаг на 1 день
	Лаг L2	Количественная	+	+	+	+	Лаг на 2 дня
	Лаг L3	Количественная	+	+	+	+	Лаг на 3 дня
	Лаг L4	Количественная	-	-	-	-	Лаг на 4 дня
	Лаг L5	Количественная	+	+	+	+	Лаг на 5 дней
	Лаг L6	Количественная	+	+	+	+	Лаг на 6 дней
	Лаг L7	Количественная	+	+	+	+	Лаг на 7 дней
Цена	Цена	Количественная	+	+	+	+	
Наличие акции	Наличие акции	Категориальная	+	+	+	+	Наличие / отсутствие акции
Наличие акции лаг	Наличие акции лаг L1	Категориальная	+	+	+	+	Лаг на 1 день
Порядковый день акции	Порядковый день акции	Количественная	-	+	+	+	Считается номер дня с начала промо-периода
Уровень скидки	Уровень скидки	Количественная	+	+	+	+	Неорицательное число
Температура воздуха	Температура воздуха	Количественная	+	+	+	+	
Порядковый день года	Порядковый день года	Количественная	+	+	+	+	Диапазон от 0 до 365/366 в зависимости от типа года

День недели	День недели	Категориальная	+	+	+	+	Указывается конкретный день недели
Наличие праздничного периода	Новогодний период	Категориальная	+	+	+	+	Наличие / отсутствие Нового года
	8 марта	Категориальная	+	+	+	+	Наличие / отсутствие 8 марта
	…	Категориальная	+	+	+	+	Наличие / отсутствие государственного праздника (учитываются предпраздничные дни)
	Пасха	Категориальная	+	+	+	+	Наличие / отсутствие Пасхи
	…	Категориальная	+	+	+	+	Наличие / отсутствие негосударственного праздника общекультурного значения (учитываются предпраздничные дни)
Порядковый день в праздничном периоде	Порядковый день в праздничном периоде	Количественная	+	+	+	+	Считается номер дня с начала праздничного периода

Страна	Страна-производитель	Категориальная	+	+	+	+	Указывается либо конкретная страна, либо группа
Производитель	Производитель	Категориальная	+	+	+	+	Указывается либо конкретная страна, либо группа
Вес (емкость) товара	Вес (емкость) товара	Количественная	+	+	+	+	
Количество чеков	Количество чеков	Количественная	+	+	+	+	
Общие продажи по группе	Общее количество продаж по группе	Количественная	+	+	+	+	
Количество товарных позиций, взаимозаменяемых по цене	Количество товарных позиций, взаимозаменяемых по цене для всех товаров	Количественная	-	+	-	-	В расчете показателя участвуют все товары
	Количество товарных позиций, взаимозаменяемых по цене для промо товаров	Количественная	+	+	+	+	В расчете показателя участвуют только промо-товары
Товарные кластеры	Товарные кластеры	Категориальная	+	+	+	+	

* - здесь и далее обозначение «+» - наличие переменной в модели, а «-» - отсутствие.

ПРИЛОЖЕНИЕ Б ПЕРЕЧЕНЬ ПЕРЕМЕННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

Общие названия переменных	Уточненное название переменной	Тип переменной	Линейная регрессия	Линейная регрессия с регуляр-й	Случайный лес	Градиентный бустинг	Примечание
Код товара	Код товара	Категориальная	-	-	-	-	Используется для анализа
Лаговая переменная спроса	Лаг L1	Количественная	+	+	+	+	Лаг на 1 день
	Лаг L2	Количественная	+	+	+	+	Лаг на 2 дня
	Лаг L3	Количественная	+	+	+	+	Лаг на 3 дня
	Лаг L4	Количественная	-	+	-	+	Лаг на 4 дня
	Лаг L5	Количественная	+	+	+	+	Лаг на 5 дней
	Лаг L6	Количественная	+	+	+	+	Лаг на 6 дней
	Лаг L7	Количественная	+	+	+	+	Лаг на 7 дней
Цена	Цена	Количественная	+	+	+	+	
	Преобразование от цен - 1/x	Количественная	-	+	-	+	Преобразование для использования особенности распределения
Наличие акции	Наличие акции	Категориальная	+	+	+	+	Наличие / отсутствие акции
Наличие акции лаг	Наличие акции лаг L1	Категориальная	+	+	+	+	Лаг на 1 день
Порядковый день акции	Порядковый день акции	Количественная	+	+	+	+	Считается номер дня с начала промо-периода
Уровень скидки	Уровень скидки	Количественная	+	+	+	+	Неорицательное число
Температура воздуха	Температура воздуха	Количественная	+	+	+	+	

Порядковый день года	Порядковый день года	Количественная	+	+	+	+	Диапазон от 0 до 365/366 в зависимости от типа года
День недели	День недели	Категориальная	+	+	+	+	Указывается конкретный день недели
Наличие праздничного периода	Новогодний период	Категориальная	+	+	+	+	Наличие / отсутствие Нового года
	8 марта		+	+	+	+	Наличие / отсутствие 8 марта
	…		+	+	+	+	Наличие / отсутствие государственного праздника (учитываются предпраздничные дни)
	Пасха		+	+	+	+	Наличие / отсутствие Пасхи
	…		+	+	+	+	Наличие / отсутствие негосударственного праздника общекультурного значения (учитываются предпраздничные дни)

Порядковый день в праздничном периоде	Порядковый день в праздничном периоде	Количественная	+	+	+	+	Считается номер дня с начала праздничного периода
Страна	Страна-производитель	Категориальная	+	+	+	+	Указывается либо конкретная страна, либо группа
Производитель	Производитель	Категориальная	+	+	+	+	Указывается либо конкретная страна, либо группа
Вес (емкость) товара	Вес (емкость) товара	Количественная	+	+	+	+	
Количество чеков	Количество чеков	Количественная	+	+	+	+	
Общие продажи по группе	Общее количество продаж по группе	Количественная	+	+	+	+	
Количество товарных позиций, взаимозаменяемых по цене	Количество товарных позиций, взаимозаменяемых по цене для всех товаров	Количественная	+	+	-	+	В расчете показателя участвуют все товары
	Количество товарных позиций, взаимозаменяемых по цене для промо товаров	Количественная	+	+	+	+	В расчете показателя участвуют только промо-товары
Товарные кластеры	Товарные кластеры	Категориальная	+	+	+	+	

ПРИЛОЖЕНИЕ В МОДУЛЬ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ НА ЯЗЫКЕ R НА ПРИМЕРЕ ПРОГНОЗИРОВАНИЯ ТЕМПЕРАТУРНОГО РЕЖИМА

```
### загрузка необходимых для моделирования пакетов
library(dplyr)
library(ggplot2)
library(prophet)
library(forecast)
library(stringi)
library(readxl)
library(lubridate)
library(bsts)
library(smooth)
library(reshape2)

### загрузка данных

load("TempData") # загрузка данных для прогнозирования температуры

#### Моделирование температурного режима
### обработка данных для моделирования температур

TempData$Date <- as.Date(TempData$Date)
TempData <- rbind(TempDataBefore, TempData)

TempDataTrain <- filter(TempData, Date < "2015-11-01")
TempDataTest <- filter(TempData, Date >= "2015-11-01")

TempDataTrainForPr <- TempDataTrain
names(TempDataTrainForPr) = c("ds", "y")

### построение модели по методу Prophet
# создание модели
modelTemp <- prophet(TempDataTrainForPr, weekly.seasonality = F,
                     seasonality.prior.scale = 20, n.changepoints = 25,
                     changepoint.prior.scale = 0.05)

# построение прогноза на будущие периоды
futureTemp <- make_future_dataframe(modelTemp, periods = nrow(TempDataTest))
forecastTemp <- predict(modelTemp, futureTemp)
plot(modelTemp, forecastTemp) # построение графика прогноза

### построение модели по методу auto.arima (алгоритм Хиндмана-Хандакара)
TempMean.ts = ts(TempDataTrain$TempMean, start = c(2008, 1), frequency = 365.25)
Koef = 5
Fu <- fourier(TempMean.ts, K = Koef) # построение рядов Фурье

# создание модели
ModelArimaFour <- auto.arima(TempMean.ts, seasonal = F, xreg = Fu)
# построение прогноза на будущие периоды
```

```
FuNew <- fourier(TempMean.ts, K = Koef, h = nrow(TempDataTest))
ModelArima.fit <- forecast(ModelArimaFour, h = nrow(TempDataTest),
                           xreg = FuNew)

### построение модели по методу ETS (модели экпоненциального сглаживания)

Koef = 1
Fu <- fourier(TempMean.ts, K = Koef)
FuNew <- fourier(TempMean.ts, K = Koef, h = nrow(TempDataTest))
FuData <- rbind(Fu, FuNew)
TempMeanGeneral.ts = ts(TempData$TempMean, start = c(2008, 1), frequency = 365.25)

# создание модели и расчет прогноза
ModelETSTemp <- es(TempMeanGeneral.ts, holdout = T, h = nrow(TempDataTest),
                   xreg = FuData, xregDo = "use")

### расчета прогноза среднего по результату

TempDataTest$ForecastProphet = forecastTempNew$yhat
TempDataTest$ForecastArima = ModelArima.fit$mean
TempDataTest$ForecastETS = ModelETSTemp$forecast
TempDataTest$Mean = rowMeans(TempDataTest[,c(3:4)])

# расчет RMSE итогового прогноза
TempDataTest$SquaredErrorMean = (TempDataTest$Mean -
                                 TempDataTest$TempMean)^2
sqrt(sum(TempDataTest$SquaredErrorMean)/nrow(TempDataTest))

# расчет MAPE итогового прогноза
TempDataTest$ErrorVarMean =
  abs((TempDataTest$Mean - TempDataTest$TempMean)/
        TempDataTest$TempMean)
sum(TempDataTest$ErrorVarMean)/nrow(TempDataTest)
```

ПРИЛОЖЕНИЕ Г МОДУЛЬ ПРОГНОЗИРОВАНИЯ СПРОСА НА ЯЗЫКЕ R

```
### загрузка необходимых для моделирования пакетов

library(dplyr)
library(caret)
library(randomForest)
library(glmnet)
library(e1071)
library(doParallel)
library(foreach)
library(bst)
library(xgboost)
library(RSNNS)
library(AUC)
library(ranger)
library(pROC)

### загрузка данных

load("DF")

### разделение выборки на обучающую и тестовую для задач классификации и регрессии

training <- filter(dfBeerModeling, Date < "2016-02-01")
testing <- filter(dfBeerModeling, Date >= "2016-02-01")

# для задачи классификации
trainingClass <- training
trainingClass <- trainingClass %>% mutate(
  Sale = as.factor(ifelse(newsales == 0, 0, 1))
)

# для задачи регрессии
trainingRegression <- training
trainingRegression <- trainingRegression %>% filter(newsales != 0)

testing <- testing %>% mutate(
  Sale = as.factor(ifelse(newsales == 0, 0, 1))
)

### моделирование оценки вероятности продажи (задача классификации)
## построение модели логистической регрессии

# моделирование стандартной функцией glm
modelGLMbase <- glm(Sale ~ (newsalesL1 + newsalesL2 + newsalesL3 +
                    newsalesL5 + newsalesL6 + newsalesL7):(cluster) +
                     weekday + YOD + numberPositionAction +
                     newprice + isaction + isactionL1 +
                     TempMean + chq_number + LevelDiscount + holiday*NDoH +
                     country + maker.Descr + weight + SalesAll,
```

```
                data = trainingClass,
                family = "binomial")

# расчет прогноза для обучающей и тестовой выборок
predLogTr = as.numeric(predict(modelGLMbase, newdata = trainingClass, type = "response"))
predLog = as.numeric(predict(modelGLMbase, newdata = testing, type = "response"))

# расчет оценки качества AUC
AUC::auc(roc(predLogTr, trainingClass$Sale))
AUC::auc(roc(predLog, testing$Sale))

## построение модели логистической регрессии с регуляризацией

levels(trainingClass$Sale) = c("NoSale", "Sale")
set.seed(141442) # фиксация значений генератора случайных чисел для воспроизводства расчетов

# параметры кросс-валидации и сетка гиперпараметров
ctrl = trainControl(method = "cv", number = 10, classProbs = T,
                    summaryFunction = twoClassSummary)
gridSet = expand.grid(alpha = c(0, 0.25, 0.5, 0.75, 1),
                      lambda = 10^seq(10, -2, length = 50))

# моделирование с помощью пакетов caret и glmnet
registerDoParallel(detectCores())# регистрация ядер процессора для реализации параллельных вычислений на CPU

modelGlmnet <- caret::train(Sale ~ (newsalesL1 + newsalesL2 + newsalesL3 +
                                      newsalesL5 + newsalesL6 + newsalesL7):(cluster) +
                      weekday + YOD + numberPosition + numberPositionAction +
                      newprice + newprice_1x + isaction + isactionL1 +
                      TempMean + chq_number + LevelDiscount + holiday*NDoH +
                      country + maker.Descr + weight + SalesAll + NDoP,
                    data = trainingClass, method = "glmnet",
                    metric = "ROC", trControl = ctrl,
                    tuneGrid = gridSet)
stopImplicitCluster()

# расчет прогноза для обучающей и тестовой выборок
predGlmnetTr = as.numeric(predict(modelGlmnet, newdata = trainingClass, type = "prob")[,2])
predGlmnet = as.numeric(predict(modelGlmnet, newdata = testing, type = "prob")[,2])

# расчет оценки качества AUC
AUC::auc(roc(predGlmnetTr, trainingClass$Sale))
AUC::auc(roc(predGlmnet, testing$Sale))

## построение модели случайного леса (без предварительного подбора гиперпараметров)

# моделирование с помощью пакета ranger
model_RF_general <- ranger(
  Sale ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL5 + newsalesL6 +
    newsalesL7 + weekday + cluster + YOD + numberPositionAction + newprice + isaction +
```

```
   isactionL1 + TempMean + chq_number + LevelDiscount + holiday + NDoH +
   country + maker.Descr + weight + SalesAll + NDoP,
  data = trainingClass, probability = T, num.trees = 500, mtry = 5,
  min.node.size = 200,
  num.threads = 8
)

# расчет прогноза для обучающей и тестовой выборок
PredRFtr = as.numeric(predict(model_RF_general, trainingClass, type = "response")$predictions[,2])
PredRF = as.numeric(predict(model_RF_general, testing, type = "response")$predictions[,2])

# расчет оценки качества AUC
AUC::auc(AUC::roc(PredRFtr, trainingClass$Sale))
AUC::auc(AUC::roc(PredRF, testing$Sale))

## построение модели градиентного бустинга

trainingClassS <- trainingClass
levels(trainingClassS$Sale) <- c("NoSale", "Sale")

# параметры кросс-валидации и сетка гиперпараметров
set.seed(491558)
trControlXGB = trainControl(method = "cv", number = 10, classProbs = T,
                            allowParallel = T)
xgb.grid <- expand.grid(nrounds = c(150,300),
                        max_depth = c(4, 6, 8),
                        eta = seq(0.02, 0.1, by = 0.02),
                        gamma = 0,
                        colsample_bytree = c(0.5, 0.7, 0.9),
                        min_child_weight = c(1,3,5),
                        subsample = c(0.5, 1))

# моделирование с помощью пакетов caret и xgboost
registerDoParallel(detectCores())
model_xgb <- caret::train(
  Sale ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL5 + newsalesL6 +
    newsalesL7 + weekday + cluster + YOD + numberPositionAction + newprice + isaction +
    isactionL1 + TempMean + chq_number + LevelDiscount + holiday + NDoH +
    country + maker.Descr + weight + SalesAll + NDoP,
  data = trainingClassS,  method="xgbTree",
  trControl=trControlXGB,
  tuneGrid=xgb.grid,
  metric = "Accuracy"
)
stopImplicitCluster()

# расчет прогноза для обучающей и тестовой выборок
predXGBtr = as.numeric(predict(model_xgb, trainingClassS, type = "prob")[,2])
predXGB = as.numeric(predict(model_xgb, testing, type = "prob")[,2])

# расчет оценки качества AUC
AUC::auc(AUC::roc(predXGBtr, trainingClassS$Sale))
```

```
AUC::auc(AUC::roc(predXGB, testing$Sale))

## комбинация результатов по оценке вероятности

predComb <- data.frame(
  Fact = testing$Sale,
  logSimple = predLog,
  logRegular = predGlmnet,
  rf = PredRF6,
  xgb = predXGB
)
corM <- cor(predComb[,-1])

# итоговый усредненный прогноз
predFinal <- rowMeans(predComb[,-c(1,3)])

# расчет оценки качества AUC
AUC::auc(AUC::roc(predFinal, testing$Sale))
AUC::auc(AUC::roc(predComb$xgb, testing$Sale))

### моделирование регрессионной оценки спроса (без учета нулевых продаж)
## построение модели линейной регрессии

# моделирование с помощью стандартной функции lm
model.base <- lm(newsales ~ (newsalesL1 + newsalesL2 + newsalesL3 +
                newsalesL5 + newsalesL6 + newsalesL7):(cluster +
                weekday + YOD + isactionL1 + newprice*isaction + SalesAll) +
                numberPositionAction + numberPosition +
                TempMean + chq_number + LevelDiscount + holiday*NDoH +
                country + maker.Descr + weight + NDoP,
                data = trainingRegression)

# расчет прогноза для обучающей и тестовой выборок
predLmTrSimple <- predict(model.base, trainingRegression)
predLmSimple <- predict(model.base, testing[testing$newsales != 0,])

# расчет оценок качества MSE и MAE
mean((predLmTrSimple - trainingRegression$newsales)^2) # MSE на обучающей выборке
mean((predLmTrSimple - trainingRegression$newsales)) # MAE на обучающей выборке
mean((predLmSimple - testing$newsales[testing$newsales != 0])^2) # MSE на тестовой выборке
mean((predLmSimple - testing$newsales[testing$newsales != 0])) # MAE на тестовой выборке

## построение модели линейной регрессии с регуляризацией

# параметры кросс-валидации и сетка гиперпараметров
set.seed(141442)
ctrlReg = trainControl(method = "cv", number = 10)
gridSetGen = expand.grid(alpha = c(0, 0.25, 0.5, 0.75, 1),
                         lambda = 10^seq(10, -2, length = 50))

# моделирование с помощью пакетов caret и glmnet
registerDoParallel(detectCores())
```

```
modelGlmnetReg <-
  caret::train(newsales ~ (newsalesL1 + newsalesL2 + newsalesL3 + newsalesL4 +
            newsalesL5 + newsalesL6 + newsalesL7):(cluster +
            weekday + YOD + isactionL1 + newprice*isaction + SalesAll) +
            numberPositionAction + numberPosition + newprice_1x +
            TempMean + chq_number + LevelDiscount + holiday*NDoH +
            country + maker.Descr + weight + NDoP,
            data = trainingRegression, method = "glmnet",
            family = "gaussian",
            trControl = ctrlReg,
            tuneGrid = gridSetGen)
stopImplicitCluster()

# расчет прогноза для обучающей и тестовой выборок
predLmTrGlmnet <- predict(modelGlmnetReg, trainingRegression)
predLmGlmnet <- predict(modelGlmnetReg, testing[testing$newsales != 0,])

# расчет оценок качества MSE и MAE
mean((predLmTrGlmnet - trainingRegression$newsales)^2)
mean((predLmTrGlmnet - trainingRegression$newsales))
mean((predLmGlmnet - testing$newsales[testing$newsales != 0])^2)
mean((predLmGlmnet - testing$newsales[testing$newsales != 0]))

## построение модели случайного леса

# параметры кросс-валидации и сетка гиперпараметров
gridSet <- expand.grid(
  .mtry = c(3, 5, 12, 18, 25, 35),
  .splitrule = "variance",
  .min.node.size = c(10, 25, 50, 75)
)

# моделирование с помощью пакетов caret и ranger
registerDoParallel(detectCores())
modelRFReg <- caret::train(
            newsales ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL5 + newsalesL6 +
            newsalesL7 + weekday + cluster + YOD + numberPositionAction +
            newprice + isaction + isactionL1 + TempMean + chq_number +
            LevelDiscount + holiday + NDoH + country + maker.Descr +
            weight + SalesAll + NDoP,
              data = trainingRegression,
              method = "ranger", importance = "impurity",
              num.trees = 500,
              trControl = ctrlReg, tuneGrid = gridSet
            )
stopImplicitCluster()

# расчет прогноза для обучающей и тестовой выборок
prRFTrReg = predict(modelRFReg, trainingRegression)
prRFReg = predict(modelRFReg, testing[testing$newsales != 0,])

# расчет оценок качества MSE и MAE
```

```
mean((prRFTrReg - trainingRegression$newsales)^2)
mean((prRFTrReg - trainingRegression$newsales))
mean((prRFReg - testing$newsales[testing$newsales != 0])^2)
mean((prRFReg - testing$newsales[testing$newsales != 0]))

## построение модели градиентного бустинга

# первая итерация
# параметры кросс-валидации и сетка гиперпараметров
set.seed(491558)
trControlXGB = trainControl(method = "cv", number = 10,
                            allowParallel = T)
xgb.gridFirstIt <- expand.grid(eta = 0.1,
                               nrounds = c(50, 100, 150, 200, 300, 400, 500),
                               max_depth = 6,
                               gamma = 0,
                               colsample_bytree = 1,
                               min_child_weight = 1,
                               subsample = 1)

# моделирование с помощью пакетов caret и xgboost
registerDoParallel(detectCores())
modelXgbReg <- caret::train(
  newsales ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL4 + newsalesL5 +
    newsalesL6 +
    newsalesL7 + weekday + cluster + YOD + numberPositionAction +
    numberPosition +
    newprice +
    isaction + newprice_1x +
    isactionL1 + TempMean + chq_number + LevelDiscount + holiday + NDoH +
    country + maker.Descr + weight + SalesAll + NDoP,
  data = trainingRegression,  method="xgbTree",
  trControl=trControlXGB,
  tuneGrid=xgb.gridFirstIt
)
stopImplicitCluster()

# вторая итерация
# параметры кросс-валидации и сетка гиперпараметров
xgb.gridSecondIt <- expand.grid(eta = 0.1,
                               nrounds = 200,
                               max_depth = c(3, 4, 6, 8, 10),
                               gamma = 0,
                               colsample_bytree = 1,
                               min_child_weight = c(1, 10, 25, 50, 100),
                               subsample = 1)
registerDoParallel(detectCores())

# моделирование с помощью пакетов caret и xgboost
modelXgbReg2 <- caret::train(
  newsales ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL4 + newsalesL5 +
    newsalesL6 +
```

```
    newsalesL7 + weekday + cluster + YOD + numberPositionAction +
    numberPosition +
    newprice +
    isaction + newprice_1x +
    isactionL1 + TempMean + chq_number + LevelDiscount + holiday + NDoH +
    country + maker.Descr + weight + SalesAll + NDoP,
  data = trainingRegression,  method="xgbTree",
  trControl=trControlXGB,
  tuneGrid=xgb.gridSecondIt
)
stopImplicitCluster()

# третья итерация
# параметры кросс-валидации и сетка гиперпараметров
xgb.gridThirdIt <- expand.grid(eta = 0.1,
                               nrounds = 200,
                               max_depth = 6,
                               gamma = 0,
                               colsample_bytree = c(0.5, 0.7, 0.9, 1),
                               min_child_weight = 10,
                               subsample = c(0.3, 0.5, 0.7, 0.9, 1))

# моделирование с помощью пакетов caret и xgboost
registerDoParallel(detectCores())
modelXgbReg3 <- caret::train(
  newsales ~ newsalesL1 + newsalesL2 + newsalesL3 + newsalesL4 + newsalesL5 +
    newsalesL6 +
    newsalesL7 + weekday + cluster + YOD + numberPositionAction +
    numberPosition +
    newprice +
    isaction + newprice_1x +
    isactionL1 + TempMean + chq_number + LevelDiscount + holiday + NDoH +
    country + maker.Descr + weight + SalesAll + NDoP,
  data = trainingRegression,  method="xgbTree",
  trControl=trControlXGB,
  tuneGrid=xgb.gridThirdIt
)
stopImplicitCluster()

# расчет прогноза для обучающей и тестовой выборок
prTrXGBreg <- predict(modelXgbReg3, trainingRegression)
prXGBreg <- predict(modelXgbReg3, testing[testing$newsales != 0,])

# расчет оценок качества MSE и MAE
mean((prTrXGBreg - trainingRegression$newsales)^2)
mean((prTrXGBreg - trainingRegression$newsales))
mean((prXGBreg - testing$newsales[testing$newsales != 0])^2)
mean((prXGBreg - testing$newsales[testing$newsales != 0]))
## комбинация результатов регрессионного моделирования

regrComb <-
  data.frame(
```

```
  Fact = testing$newsales[testing$newsales != 0],
  lmSimple = predLmSimple,
  lmReg = predLmGlmnet,
  rf = prRFReg,
  xgb = prXGBreg
 )

corR <- cor(regrComb[,-1]) # матрица корреляций для оценки зависимости результатов

# итоговый усредненный прогноз
regFinal <- rowMeans(regrComb[,-c(1,2)])

# расчет оценок качества MSE и MAE
mean((regFinal - testing$newsales[testing$newsales != 0])^2)
mean((regFinal - testing$newsales[testing$newsales != 0]))

### итоговая комбинация результатов задач классификации и регрессии (прогноз спроса)

predLmGlmnetFinal = predict(modelGlmnetReg, testing)
prRFRegFinal = predict(modelRFReg1, testing)
prXGBregFinal = predict(modelXgbReg3, testing)

## значение прогноза спроса вариант 1 (в основе градиентный бустинг)

FinalY_1 = predComb$xgb*((predLmGlmnetFinal+prRFRegFinal+prXGBregFinal)/3)
mean((FinalY_1 - testing$newsales)^2)
mean((FinalY_1 - testing$newsales))

## значение прогноза спроса вариант 1 (в основе логистическая регрессия) - финальный

FinalY_2 = predComb$logSimple*((predLmGlmnetFinal+prRFRegFinal+prXGBregFinal)/3)
mean((FinalY_2 - testing$newsales)^2)
mean((FinalY_2 - testing$newsales))
```

ПРИЛОЖЕНИЕ Д АКТ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

«УТВЕРЖДАЮ»
ООО «Гастроном»
_____ О. В. Бокова
«18» мая 2018 г.

А К Т

о внедрении результатов диссертационной работы «Моделирование покупательского спроса на предприятиях розничной торговли на основе методов машинного обучения» автора Пивкина Кирилла Сергеевича

Мы, представители ООО «Гастроном», настоящим актом подтверждаем, что результаты ряда научно-исследовательских и организационно-методических работ, выполненных в рамках проекта развития компании «Автозаявка», направленных на повышение эффективности системы управления товарными запасами и улучшения качества прогнозирования внутренних процессов предприятия, были получены при непосредственном участии бывшего бизнес-аналитика ООО «Гастроном» и соискателя на научную степень кандидата экономических наук Пивкина Кирилла Сергеевича, который внес существенный творческий и научный вклад в решение проблемы совершенствования системы управления товарными запасами Компании.

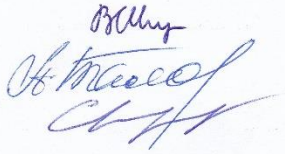
Практическая ценность диссертационной работы заключается в том, что предложенная система математического моделирования спроса позволяет оптимизировать уровень товарных запасов предприятия при непосредственном ее подключении к системам заказа. В свою очередь полученная система моделирования выражена в программном обеспечении на языке программирования R, которая легко интегрируется в информационные системы предприятия. Фактически, это позволяет оптимизировать товарный запас, увеличить производительность труда работников, ответственных за заказ товара, повысить лояльность клиентов к бренду «Гастроном» за счет сокращения случаев неудовлетворенного спроса. В итоге это ведет к увеличению прибыли Компании.

Научно-экономическая новизна полученных результатов диссертационной работы состоит в построении экономико-математической модели прогнозирования спроса с учетом покупательских реакций на изменение ценовой политики, насыщенности ассортимента и внешних эффектов (температурный режим), в разработке сопутствующих переменных для итогового набора данных, основанных на гипотезах о поведении

покупателей в магазине как экономических агентов, в построении моделей прогнозирования ключевых показателей торговой компании на основе ансамбля временных рядов, использующихся как самостоятельно в планировании на предприятии (в частности в планировании товарооборота), так и в составе модели прогнозирования спроса, в создании автоматизированного сервиса прогнозирования спроса, который учитывает иерархию открытых магазинов торговой сети и ее товарных групп. В основе моделирования стоят инструменты эконометрического анализа и машинного обучения, которые актуальны для современных научных разработок как в области розничной торговли, так и в области информационных технологий в целом.

Полученные в ходе диссертационного исследования результаты были использованы в процессе управления товарными запасами, наибольших результатов удалось добиться в формализованном товаре (штучный товар).

Применяя разработанные экономико-математические модели, ООО «Гастроном» за период с 01.01.2017 по 31.12.2017 снизил средние товарные запасы на 9,6% и при этом увеличил продажи товара на 4,3%.

Зам. директора по ИТ		С.В. Широбоков
Коммерческий директор		А.А. Балобанов
Инженер-программист		С.С. Королев

ПРИЛОЖЕНИЕ Е АКТ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

УТВЕРЖДАЮ
Проректор по научной работе и
программам стратегического развития
Удмуртского государственного
университета,
доктор экономических наук, профессор

_____ Макаров А.М.

«___» _____ 2018 г.

АКТ
о внедрении результатов диссертационной работы
на соискание ученой степени кандидата экономических наук
Пивкина Кирилла Сергеевича

Результаты диссертационной работы Пивкина Кирилла Сергеевича «Моделирование покупательского спроса на предприятиях розничной торговли на основе методов машинного обучения» на соискание ученой степени кандидата экономических наук включены в учебный комплекс дисциплин «Эконометрическое моделирование», «Моделирование бизнес-процессов», «Информационные системы управления производственной компанией».

Заведующий кафедрой
финансов, учета и математических
методов в экономике, кандидат
эконом. наук, доцент _____ Федулова С.Ф.