УДК 330.43+ 519.874

К.С. Пивкин

АЛГОРИТМ ПОСТРОЕНИЯ ЛИНЕЙНОЙ МОДЕЛИ НА ПАНЕЛЬНЫХ ДАННЫХ КАК ЭТАП ЭКОНОМЕТРИЧЕСКОГО ПРОГНОЗИРОВАНИЯ ТОВАРНОГО СПРОСА

Рассматривается роль прогнозирования спроса в автоматизации процессов в розничной торговле. Приводится описание возможных решений задачи прогнозирования в виде подзадачи регрессионного анализа. Описывается эвристический подход к формированию данных для оценки модели множественной линейной регрессии на целевую переменную – покупательский спрос на товар. Выводятся новые независимые переменные, которые имеют обоснованный экономический характер. Приводится последовательная аргументация о содержании и типе независимых переменных, участвующих в эконометрическом моделировании спроса. Разрабатывается базовая модель прогнозирования с описанием основных характеристик, коэффициентов и метрики качества модели. Подтверждается ряд гипотез, сделанных о типе связи независимых переменных с целевой переменной. Подчеркивается характер результата как промежуточного в моделировании покупательского спроса на товар с применением методов параметрического и непараметрического регрессионного анализа. Делаются выводы о результате работы и векторе развития дальнейшего эконометрического исследования. Предлагается разработать новые модели на основании полученных данных, а также разработать их композицию (ансамбль). Рассматривается язык R не только как инструмент статистического анализа, но и как среда разработки продвинутой модели прогнозирования спроса.

Ключевые слова: эконометрическое прогнозирование, розничная торговля, покупательский спрос на товар, панельные данные, статистические методы, множественная линейная регрессия, язык программирования R.

Введение

Этап развития российской экономики, мировых управленческих и информационных технологий говорит о смещении вектора с простого учета, накопления экономической информации и простых арифметических операций над ней на развитие сложных вычислительных обработок с целью автоматизировать розничный бизнес. Весь решаемый спектр задач сложно перечислить: создание рекомендательных сервисов, где продукт будет предлагаться возможному покупателю автоматически; автоматизированное создание планов торговых площадей и планограмм с размещаемыми на нем товарными группами / позициями (решение проблем мерчандайзинга); оптимизация выкладки товара: периодичности, объема выкладки и затрачиваемого времени на данную работу ответственным менеджером; прогнозирование товарного спроса по каждой группе товаров и конкретно по товарным позициям и т. п. Цель этих задач – автоматизация бизнес-процессов розничной компании, то есть сокращение издержек на выполнение подобной работы наемными специалистами. При это речь идет не только о сохранении качества работы, но и о его повышении ввиду ограничений человеческого мышления и существования так называемого «человеческого фактора» [10].

В данной статье будет рассматриваться одна из ключевых задач в области как современной розничной торговли, так и сферы «data mining» и эконометрического моделирования – это прогнозирование товарного спроса по конкретной товарной позиции (SKU – *Stock Keeping Unit*). Предназначение прогнозирования товарного спроса по SKU заключается в усовершенствовании управления товарными запасами, оптимизации трудозатрат и разработке тактических планов по продвижению товаров. Таким образом, прогноз спроса является базисом для дальнейшего вектора развития розничной компании, его качество задает эффективность бизнес-процессов компании.

Для качественного решения задачи прогнозирования необходимы не только большие объемы накапливаемых данных, но и специальные знания в области математического моделирования и статистической обработки информации, использование специализированных программных комплексов и языков программирования. Обычно прогнозирование товарного спроса рассматривают как прогнозирование некоторого количественного отклика с учетом специфики задачи и рассматриваемых данных, а также функционала качества модели [4; 12]. На данный момент имеется огромное количество математических методов и моделей, которые по-своему решают данную задачу: это и методы регрессионного анализа – разные виды линейной регрессии, деревья решений, нейронные сети и т. п. [8; 15]; также методы анализа временных рядов – модели экспоненциального сглаживания, авторег-

рессионные модели из разновидности и сочетания [1]. В целом ясно, что просто использование каких-либо методов не дает должного результата. Для успешного прогнозирования необходимо, во-первых, подготовить исходные данные, которые отражают логику процессов, происходящих при продаже определенного продукта, во-вторых, протестировать определенные методы и выявить лучшие с точки зрения функционала качества, в-третьих, создать ансамбль моделей [2] или определенные адаптивные механизмы сочетания моделей для того чтобы нивелировать риск нарастания ошибки прогноза с течением времени [13].

В данной статье рассматриваются этапы выбора и процесс предобработки данных, разработка базовой математической модели и сравнительный анализ ее результатов с существующей на предприятии розничной торговли простой модели прогнозирования. В данных учитывается логика выбора покупателем товара из всего спектра многономенклатурного предложения и тем самым улучшается результат прогнозирования спроса (хороший пример по логике принятия решений покупателем можно найти в [11]).

1. Выбор исходных данных и их предобработка

Согласно исследованию, проведенному в статье [7], основными внешними и внутренними факторами, влияющими на товарный спрос, являются:

- внутренняя динамика (характеристика) продаж товара;
- активность покупательского потока в торговой точке;
- наличие календарного праздника и прочие календарные эффекты (день недели).

Производится выборка по всем товарам конкретной группы заданной торговой точки сети розничных магазинов. Форма используемых в анализе данных представляется в виде «панельных данных» с наличием целевой переменной в виде продаж конкретного товара, указания метки (кода) товара и прочих зависимых переменных. Период сформированной выборки – минимум 2,5 года истории продаж; в использованных для вывода результатов конкретных данных период задан с 01.10.2013 по 30.09.2016 г. Здесь необходимо отметить, что период обусловлен наличием определенной годовой сезонности, которая должна быть зафиксирована в данных для более качественного прогноза.

Перед описанием логики данных требуется обозначить общие принципы, по которым строятся первоначальные предположения о данных, а также их предобработка:

- первоначальные гипотезы о структуре и составе данных основаны на исследовании в статье [7] и следующих источниках [3; 5; 6];
- по типу переменные стандартно разделяются на количественные, номинальные и порядковые. В ходе анализа характер первоначальных переменных изменялся, исходя из определенных предпосылок и критерия качества базовой модели;
- базовой моделью в рамках определения спецификации и природы данных и их взаимосвязей является множественная линейная регрессия общего вида:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,$$

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 может меняться от заданной аддитивной и, в зависимости от качества модели, может сочетать в себе также нелинейные зависимости. Изменения в рассматриваемой спецификации будут основаны на методе эвристического поиска минимума такого критерия качества модели, как средняя квадратическая ошибка (MSE):

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

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

• обучающая и тестовая выборки формируются из исходной в пропорциях 80% и 20% соответственно, строго на 2 временных интервалах. Исследователем выстраиваются первичные гипотезы о составе и типе данных, рассчитываются коэффициенты модели при данном наборе данных, затем оценивается коэффициент MSE на тестовой выборке. MSE базовой модели сравнивается с аналогичным значением, полученным с помощью существующего на предприятии розничной торговли алгоритма прогнозирования:

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

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. Метод, по сути, является алгоритмом взвешенной скользящей средней.

Дополнение и преобразование переменных производится до того момента, пока

$$MSE_{lm} < MSE_{MA},$$

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;

• задача исследователя состоит в том, чтобы сформировать переменные, которые наиболее полно отражают анализируемые процессы, и разработать базовую модель с категорией качества выше качества текущей модели.

Данную процедуру можно представить с помощью следующей блок-схемы:

Рис. 1. Алгоритм эвристического поиска и преобразования независимых переменных

Далее приводится описание всех использованных переменных, процедура обработки и логика их использования при создании модели прогнозирования спроса.

Код (идентификатор) и наименование товара. Единица товарной номенклатуры определяется понятием Stock Keeping Unit (SKU) или идентификатор товарной позиции. У каждого SKU свой уникальный код и уникальное наименование. В сформированной выборке содержатся данные по 691 уникальным товарным позициям, с собственной историей продаж и жизненным циклом. Корректное нахождение взаимосвязей между товарами позволит улучшить качество прогноза.

Дата (период). Единица времени (периода), подходящего для решения сформулированной задачи, – это день. Соответственно выборка сформирована по каждому дню с 01.10.2013 по 30.09.2016 г. Выбор в качестве периода дневных срезов обусловлено вопросом пополнения и управления товарными запасами. На практике подавляющее большинство поставщиков розничной сети имеют график поставки товара не чаще 1 раза в день. Таким образом, 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 следующим образом:

• обучающие данные: по периоду с 01.10.2013 по 01.02.2016, что составляет примерно 79% от исходной выборки;

• тестовые данные: по периоду – с 01.02.2016 по 30.09.2016, что составляет 21% от исходной выборки.

Четкая привязка разделения выборок к времени обусловлена тем, что задача связана с прогнозом значений спроса на будущие периоды. Соответственно, на прогноз может влиять наличие нестационарности, 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 предпразднич-

ных / 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 (подробно в [7]).

Остатки товара на начало и на конец дня. Переменные, которые выражают количество товара, находящегося 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. Является первообразной переменной по отношению к моделируемой переменной товарного спроса. Для выявления из продаж товара непосредственно значений *спроса* необходимо выполнить следующие условия:

$$y_i = \begin{cases} NA, если\ b_i \le 0 \\ NA, если\ e_i \le 0, \\ иначе\ s_i \end{cases}$$

где y_i – величина покупательского спроса на товар, b_i – значение величины остатка товара на начало дня, взятое из информационной базы предприятия, e_i – значение величины остатка товара на конец дня, взятое из информационной базы предприятия, s_i – значение величины продаж товара, NA – отсутствующее значение целевой переменной [3].

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. Для представления полученной переменной спроса y выводится гистограмма:

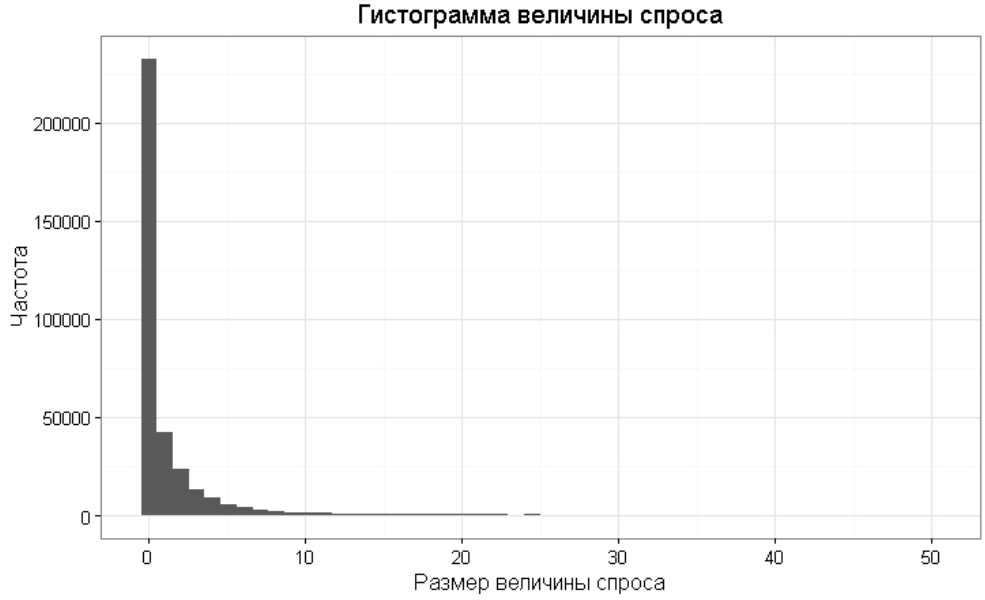

Рис. 2. Распределение величины спроса

Видна концентрация значения в пространстве около 0, что только подтверждает факт принятия в качестве функционала качества MSE. Это позволит лучше учесть качество прогноза продаж в правом хвосте распределения остатков $\left(Y - \hat{Y}\right)^2$.

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 [1; 7]. Тем не менее выбор включаемых в модель лагов является нетривиальной задачей с точки зрения конечного качества модели.

В качестве примера приведем график частной автокорреляционной функции одного из товаров (выбран по SKU) исходной выборки:

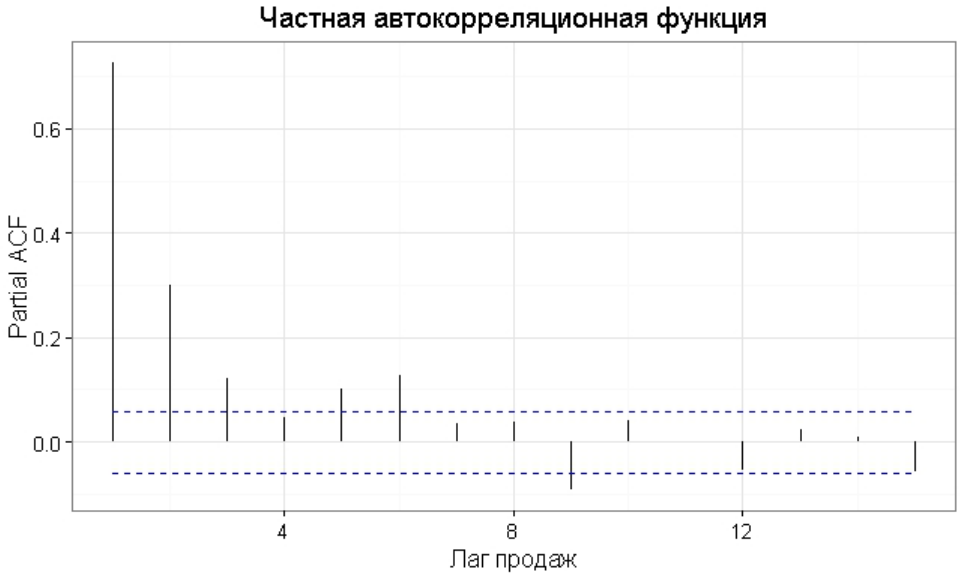

Рис. 3. Частная автокорреляционная функция продаж товара

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 будет выглядеть следующим образом:

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

где y – целевая (зависимая) переменная, $y_{t-1}, y_{t-2}, \ldots, y_{t-k}$ – лагированные значения ряда, $\alpha_1, \alpha_2, \ldots, \alpha_k$ – 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. Состав включаемых лагируемых переменных может меняться в зависимости от условия максимизации качества модели.

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

$$l = \left| \frac{c_b - c_p}{c_b} \right|,$$

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

Наличие ценовой акции. Факторная переменная, которая выражает в себе наличие ценовой акции на товар со всеми сопутствующими 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 добавим в исходную выборку лагированную переменную «Наличие ценовой акции», которая позволит выравнить 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, 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» [9; 14].

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 [8]. Данная процедура повышает предсказательную способность и устойчивость модели на тестовой выборке.

Количество SKU, взаимозаменяемых по цене. Количество товарных позиций, взаимозаменяемых по цене – это показатель, характеризующий распределение ассортимента в зависимости от категории цены по данной товарной группе. Он определяется следующим образом:

$$N_{SKU} = \sum N \mid P \in (pr, pr + h],$$

где N_{SKU} – количество SKU, взаимозаменяемых по цене, N – пространство уникальных товарных единиц, ограниченное пространством цен P, pr – конкретная цена товара (нижняя граница интервала), h - шаг цены, формируемый, исходя из следующего принципа:

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

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 доли, возможные значения которого принимаются от 0 до 1. Значение pr_α используется при определении шага для объединения экстремально высоких значений цен в группу «pr_α и выше». Исходя из постановки, разбиение показателя N_{SKU} зависит от параметра d. Параметр d выбирается руководствуясь соображениями о товарной заменимости в группе.

Показатель N_{SKU} вводится в исходную выборку для отражения предположения о том, что товары с ценой в рамках одного диапазона при прочих равных условиях имеют показатель средних продаж, стремящийся к 0, при условии:

$$\bar{s} = \lim_{N_{SKU} \to +\infty} \frac{S}{N_{SKU}} \to 0,$$

где S – суммарные продажи товаров (емкость) в рамках одного ценового диапазона $(pr_i, pr_i + h]$, N_{SKU} – количество товарных позиций внутри заданного ценового диапазона, $\bar{s}$ – средние продажи товаров на 1 наименование внутри заданного ценового диапазона.

Предположение основано на том, что у каждого ценового диапазона группы товаров имеется свой покупатель с определенным набором характеристик по доходу и определенным бюджетом на покупку той или иной группы товаров. Соответственно, если ассортиментный состав в группе увеличивается, то у покупателя безразличного к бренду, производителю и др. факторам появляется большая возможность выбора товара на тот же выделяемый бюджет и средняя покупка на одно наименование уменьшается; если количество SKU в ценовой категории уменьшается, то наоборот – идет увеличение средних продаж. Следовательно, если устремить N_{SKU} к $+\infty$, то средние продажи на 1 наименование на каком-то заданном горизонте усреднения будут стремиться к 0 из-за возникновения огромного количества выбора товаров внутри диапазона. Косвенно данное предположение подтвердится в случае наличия отрицательного коэффициента для переменной N_{SKU} в модели множественной линейной регрессии.

Товарные кластеры. Основным преимуществом при моделировании целевой переменной на панельных данных является реализация системного подхода к рассматриваемой задаче. Тем не менее существует определенного рода проблема – при таком подходе в качестве товарных идентификаторов при моделировании спроса невозможно использовать код или наименование товара. Это обусловлено тем, что при реализации процедуры обучения модели данные разобьются на достаточно мелкие выборки, равные по количеству размерности пространства фиктивных переменных $X_{Cd_{p-1}}$, где p – это количество уникальных кодов товаров. В результате этого разбиения, моделирование целевой переменной будет происходить на малых данных, что приведет к необученной модели и большой ошибке на тестовой выборке. Соответственно, в целях моделирования спроса необходимо выделить определенные товарные кластеры, которые будут объединять схожие по характеристикам товары в определенные группы.

Делается предположение, что товар определяется пространством следующих характеристик:

- вес (емкость) товара;
- цена товара в определенный момент времени;
- наличие или отсутствие ценовой акции по товару;
- принадлежность товара к конкретному производителю (пространство фиктивных переменных $X'_{M_{k-1}}$ после удаления не вариативных переменных);
- принадлежность товара к конкретной стране (пространство фиктивных переменных $X'_{C_{m-1}}$ после удаления не вариативных переменных).

Необходимо сгруппировать товары, исходя из этих характеристик. Наиболее подходящим методом считается метод кластерного анализа *k-means* (метод k-средних).

Для поднастройки количества кластеров используется значение MSE линейной модели на тестовой выборке. По результатам настройки количество кластеров остановилось на 13 – с данным количеством кластеров MSE на тестовой выборке минимально. По характеристике полученных центров кластеров есть возможность провести содержательный анализ полученных групп.

2. Базовая линейная модель: использование в формировании данных

На этапе алгоритма на рис. 1 после определения всех переменных и их размерностей проводится оценка коэффициентов множественной линейной регрессии с аддитивным эффектом. В табл. 1 частично приведена оценка коэффициентов модели (всего в модели 60 переменных):

Таблица 1

Коэффициенты модели линейной регрессии (без нелинейных преобразований)

Показатель	Оценка коэффициента	Станд. 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

Окончание табл. 1

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 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.

Таблица 2

Коэффициенты модели линейной регрессии (с нелинейными преобразованиями переменных)

Показатель	Оценка коэффициента	Станд. 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 производители	0,270	0,024	11,480	0,000
Производитель 1	–0,055	0,026	–2,092	0,036
…	…	…	…	…
Емкость (вес)	–0,027	0,008	–3,545	0,000

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. 1): $MSE_{lm} > MSE_{MA}$, где $MSE_{MA} = 10{,}55$. Для улучшения результата в рамках модели необходимо произвести дальнейшие преобразования формулы линейной регрессии. В ходе преобразований также повторяется алгоритм, описанный на рис. 1.

Итоговая модель интерпретируется следующей формулой:

$$y = \beta_0 + (\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_k y_{t-k}) \times (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{m-l} x_{m-l}) + \cdots + \beta_m x_m + \varepsilon,$$

где y – целевая (зависимая) переменная, $y_{t-1}, y_{t-2}, \ldots, y_{t-k}$ – лагированные значения ряда, $\alpha_1, \alpha_2, \ldots, \alpha_k$ – 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 несколько изменилась – существует 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: $MSE_{lm} = 9{,}09$, что является меньшим значением, чем MSE_{MA}.

Параметры оценки некоторых из рассматриваемых переменных сведены в табл. 2.

Полученная модель имеет объясненный $R^2 = 0{,}8038$, MSE_{lm} на тестовой выборке равен 9,09. Следовательно, в соответствие с заданным алгоритмом выбор исходных переменных и их преобразованных вариантов завершается.

Выводы

Видно, что применение эвристического алгоритма, заданного на рис. 1, позволяет исследователю достичь основного результата:

1) определить основные переменные, извлеченные из исходных данных информационной системы, а также кардинально новые преобразованные данные, которые позволяют учесть специфику принятия решений покупателем;

2) создать базу для альтернативы текущему методу, основанному на взвешенной скользящей средней, которая обладает недостаточной прогностической способностью.

Все же стоит отметить ряд вопросов к рассматриваемому алгоритму для улучшения качества исполнения задачи по прогнозированию товарного спроса:

- метод, как и многие другие, требует предварительного анализа перемененных включения: экономического, статистического и других. В рамках данной статьи подробный анализ не приводится, тем не менее с его методология приведена в источниках, указанных выше [3; 5-7];
- результатом моделирования является линейная модель с большим количеством коэффициентов. Дальнейшими шагами в развитии решения задачи прогнозирования спроса является применение более продвинутых моделей регрессионного анализа, где снижается объясняющая способность модели, но повышается ее точность. Также следует рассмотреть возможность композиции моделей;
- существует ряд переменных, например, температурный режим или количество чеков, которые также имеют случайный характер, но включены в базовую модель. Поэтому для них требуется разработка отдельных алгоритмов прогнозирования, которые будут основаны на анализе временных рядов (ARIMA, тренд-сезонные и иные модели);
- не смотря на существование «мягкого» условия конечности алгоритма (на рис. 1), вполне возможны дальнейшие концептуальные изменения в данных и обогащение исходной выборки;
- в качестве технологической базы для дальнейшего моделирования предлагается использовать язык программирования R с пакетными расширениями для работы над продвинутыми методами – машиной опорных векторов, случайным лесом и искусственными нейронными сетями.

СПИСОК ЛИТЕРАТУРЫ

1. Учебник СтатСофт по статистике. Раздел: Анализ временных рядов. URL: http://statsoft.ru/home/textbook/default.htm.
2. Business Data Analytics: Ансамбли моделей. URL: http://businessdataanalytics.ru/ModelEnsembles.htm.

3. Баль А.В., Логиновский О.В. Автоматизированный заказ высокооборачиваемых товаров с низкими сроками годности с использованием почасовых продаж // Вестн. Южно-Уральского гос. ун-та. Сер. Компьютерные технологии, управление, радиоэлектроника. 2015. Т. 15, Вып. 1. С. 21-25.
4. Баринова О.В., Вальков А.С., Воронцов К.В., Громов С.А., Ефимов А.Н., Чехович Ю.В. Система прогнозирования потребительского спроса Goods4Cast. Вычислительный центр им. А.А. Дородницына РАН. М., 2015.
5. Винжегин О.М. Прогнозирование продаж товаров, обладающих сезонным спросом (на примере пива) // Вестн. Омского ун-та. Серия «Экономика». 2006. № 3. С. 128-130.
6. Джалилова У.Т. Моделирование спроса на продовольственные товары на основе учета его особенностей // Вестн. Таджик. гос. ун-та права, бизнеса и политики. Серия гуманитарных наук. 2014. № 1 (57). С. 159-164.
7. Пивкин К.С. Корреляционный анализ факторов влияния на покупательский спрос розничного магазина как этап формирования модели прогнозирования и управления запасами // Вестн. Удм. ун-та. Сер. Экономика и право. 2016. Вып. 3. С. 40-50.
8. Kuhn M., Johnson K. Applied Predictive Modeling. Springer, 2013. 600 p. 203 illus., 153 illus. in color.
9. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 1: в 2 кн. М.: Финансы и статистика, 1986. 366 с.
10. Кристеноен Ж., Мейстер Д., Фоули П. и др. (Gavriel Salvendy). Человеческий фактор: в 6 т. Т. 1: Эргономика – комплексная научно-техническая дисциплина: = Handbook of Human Factors / В.П. Зинченко, В.М. Мунипов. М.: Мир, 1991.
11. Крок Г.Г., Сысоева С.В. Большая книга директора магазина 2.0. Новые технологии. СПб.: Питер, 2016. 464 с.
12. Линдерс М.Р., Фирон Х.Е. Управление снабжением и запасами. Логистика: зарубежный учебник / пер. с англ. М.: Юнити-Дана, 2007. 752 с.
13. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003. 416 с.
14. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: учебник. 6-е изд., перераб. и доп. М.: Дело, 2004. 576 с.
15. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. 60 с.

Поступила в редакцию 26.02.17

K.S. Pivkin

ALGORITHM OF BUILDING A LINEAR MODEL USING PANEL DATA AS A STAGE OF ECONOMETRIC FORECASTING OF DEMAND FOR GOODS

The role of demand forecasting in process automation in the retail trade is considered. A description of possible solutions to the problem of forecasting in the form of regression analysis subtasks is given. The heuristic approach to building data to estimate the basic model of multiple linear regression on the target variable – consumer demand for goods is described. New independent variables that have a reasonable economic nature are derived. Consistent arguments are offered concerning the content and type of independent variables used in the econometric modeling of demand. A basic prediction model is developed which involves a description of the main characteristics, coefficients and the metric of the quality of the model. A number of hypotheses about the type of connection of the independent variables with the target variable are confirmed. An intermediate nature of the result in the modeling of consumer demand for goods is emphasized by using methods of parametric and nonparametric regression analysis. Conclusions are drawn about the results of the study and about the vector of development of further econometric research. It is proposed to develop a new model based on the data obtained and to develop their composition (ensemble). The language R is considered not only as a tool for statistical analysis, but also as a development environment for an advanced demand forecasting model.

Keywords: econometric forecasting, retail, customer demand for goods, panel data, statistical methods, multiple linear regression, programming language R.

Пивкин Кирилл Сергеевич, аспирант

ФГБОУ ВО «Удмуртский государственный университет»
426034, Россия, г. Ижевск, ул. Университетская, 1 (корп. 1)
E-mail: cmme@uni.udm.ru

Pivkin K.S., postgraduate student

Udmurt State University
Universitetskaya st., 1/1, Izhevsk, Russia, 426034
E-mail: cmme@uni.udm.ru