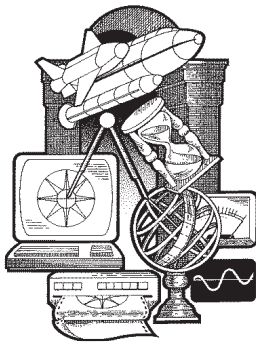

С. В. ГАВРИЛОВ

ПРЕИМУЩЕСТВО МАТЕМАТИЧЕСКИХ ИНСТРУМЕНТОВ ПРОГНОЗИРОВАНИЯ

В статье анализируются принципы работы ряда алгоритмов машинного обучения, описываются сценарии их применения для решения реальных задач прогнозирования на примере сферы розничной торговли. Результатом исследования является подтверждение возможности внедрения машинного обучения в практику построения прогнозов и подтверждение высокой эффективности алгоритмов.

Ключевые слова: машинное обучение; прогнозирование продаж; управление запасами.

УДК 338.984

Введение. Прогнозирование объемов продаж традиционно играет ключевую роль в управлении запасами и их оптимизации и непосредственно влияет на эффективность цепи поставок. Точность прогнозов позволяет минимизировать избыточность запасов, эффективно планировать производство и в конечном итоге — максимизировать прибыль. Обратной стороной такой важности прогнозов является чрезвычайно высокая цена ошибки. На фоне постоянного роста спроса и емкости рынков становится все труднее учитывать многочисленные факторы, влияющие на поведение потребителя, поскольку каждый из факторов оказывает уникальное воздействие на спрос и не всегда может быть оценен достаточно точно, а главное — достаточно своевременно, чтобы динамически адаптировать прогнозы при возникновении соответствующих предпосылок.

Традиционные техники прогнозирования, такие как скользящее среднее, экспоненциальное сглаживание, модель Хольта — Уинтерса, были разработаны и внедрены в практику построения прогнозов в цепях поставок еще в прошлом столетии. Принципиальное различие в способах их применения в сравнении с тем периодом заключается лишь в автоматизации вычислений с помощью современного программного обеспечения. В настоящее время они по-прежнему широко используются и остаются самым распространенным способом построения прогнозов.

Сергей Владимирович ГАВРИЛОВ (gavrilovsv_95@mail.ru)*, аспирант кафедры маркетинга Белорусского государственного экономического университета (г. Минск, Беларусь).*

Традиционные методы используются при решении следующих задач:

– построение средне- и долгосрочных прогнозов;

– прогнозирование объемов продаж продуктов со стабильным спросом, как правило, давно присутствующих на рынке;

– прогнозирование суммарных объемов продаж больших товарных групп и предприятия в целом [1].

Тем не менее современные рыночные тенденции характеризуются сокращением жизненного цикла продукта и повышением волатильности спроса. Кроме того, возрастает объем информации, доступной для анализа, и становится актуальной задача поиска неочевидных зависимостей среди огромного массива доступных данных. В связи с этим растет интерес к новым инструментам, способным решать задачи прогнозирования с более высокой точностью, минимизируя тем самым риски, возникающие при ошибках прогнозирования.

Целью статьи является изучение возможности применения современных алгоритмов машинного обучения для решения задач экономического прогнозирования, в частности — прогнозирования объемов продаж.

Основная часть. *Машинное обучение* (Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Иначе говоря, машинное обучение предполагает создание системы, способной самостоятельно решать задачи определенного класса, например, в области прогнозирования объемов продаж.

Алгоритмы машинного обучения нельзя назвать новыми — их математическая основа была сформулирована еще в 50—60-х гг. прошлого столетия. Однако эта область долгое время была не востребована в силу отсутствия достаточных мощностей для реализации столь сложных вычислительных процедур.

В основе алгоритмов лежат статистические модели, а сами алгоритмы базируются на использовании всевозможных источников информации, как внутренней (для цепи поставок), так и внешней (маркетинговой, макроэкономической, социальной и любой другой), поддающейся числовой оценке — вплоть до прогнозов погоды и температуры окружающей среды.

Машинное обучение предполагает применение сложных математических алгоритмов для автоматического отслеживания поступающих сигналов и выявления неочевидных взаимосвязей в больших объемах данных. Помимо непосредственной возможности глубокого анализа данных преимущество алгоритмов состоит еще и в возможности непрерывного самосовершенствования систем на их основе и их своевременной адаптации к малейшим изменениям внутренних и внешних факторов. Все это в совокупности позволяет получать более точные и надежные прогнозы в сложных сценариях.

Недостатком подхода является отсутствие прозрачности решений, принимаемых в рамках алгоритма. По сути, система, построенная на основе машинного обучения, действует по принципу черного ящика, где внутренние механизмы скрыты и не подлежат интерпретации, а оптимизация сводится лишь к возможности оперирования параметрами самой модели и набором данных для обучения.

С учетом описанных преимуществ и недостатков, применительно к задаче прогнозирования, алгоритмы машинного обучения наилучшим образом подходят в следующих случаях:

– построение кратко- и среднесрочных прогнозов;

– прогнозирование для рынков с высокой волатильностью спроса;

– построение прогнозов в быстро изменяющихся условиях;

– прогнозирование для новых продуктов [1].

В статье будут рассмотрены две модели машинного обучения, основанные на следующих алгоритмах: Random Forest и Gradient Boosting.

В основе этих алгоритмов лежит понятие «ансамбль» (Ensemble), под которым понимается набор отдельных прогнозов, используемых совместно для получения итогового результата. Смысл использования ансамблей заключается в том, что комбинация множества независимых решений в большинстве случаев показывает лучшие результаты, чем отдельно взятое решение.

Техники ансамблирования подразделяются на две группы — Бэггинг (Bagging) и Бустинг (Boosting).

Бэггинг представляет собой простую технику ансамблирования, при которой формируется множество независимых результатов, на завершающем этапе комбинируемое в итоговый результат, например с помощью усреднения. Примером подобного алгоритма является Random Forest.

В основе алгоритма Random Forest («случайный лес») лежит понятие «дерево решений». Дерево решений — это метод представления решающих правил в иерархической структуре, состоящей из элементов двух типов — узлов и листьев. В узлах располагаются решающие правила и производится проверка соответствия примеров этому правилу по какому-либо атрибуту обучающего множества.

В простейшем случае, в результате проверки множество примеров, попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое — не удовлетворяющие.

Затем к каждому подмножеству вновь применяется правило, и процедура рекурсивно повторяется, пока не будет достигнуто некоторое условие остановки алгоритма. В результате в последнем узле проверка и разбиение не производится, и он объявляется листом. Лист определяет решение для каждого попавшего в него примера.

Суть алгоритма Random Forest заключается в создании множества подобных деревьев, каждое из которых обучается принятию решений на основе анализа *случайного* набора факторов из всего множества доступных факторов. В итоге каждое из деревьев принимает решение об итоговом значении выходной переменной, после чего результаты отдельных деревьев комбинируются и на выходе получается окончательный результат [2, с. *353*].

Суть алгоритма хорошо передает аналогия с прогнозированием объема продаж путем опроса группы экспертов и последующего нахождения среднего значения их прогнозов. Каждый из экспертов будет делать прогноз на основании того набора факторов, который он считает наиболее существенным, исходя из собственного опыта. Поскольку алгоритм не может иметь собственного опыта, который позволил бы ему принимать решения, необходим процесс обучения, в результате которого каждое отдельно взятое дерево решений (по аналогии с отдельно взятым экспертом) учится сопоставлять значение набора случайных факторов с итоговым результатом.

Техника Бустинга принципиально отличается от Бэггинга — в этом случае отдельные прогнозы строятся не независимо, а последовательно. Алгоритмы, использующие эту технику, основываются на обучении на основании ошибок предыдущих шагов. Эти алгоритмы являются итеративными и предполагают расчет некоторого набора весов на каждой итерации таким образом, чтобы неверно обработанные входные данные получили больший вес, а корректно обработанные — меньший. Тем самым следующая итерация фокусируется на примерах, где предыдущая итерация допустила ошибки. В результате строится комбинация последовательных элементов, каждый из которых может принимать простые решения (к примеру, в качестве простых элементов могут использоваться те же бинарные деревья решений) таким образом, чтобы на выходе получить высокоточную модель [3].

Примером подобного алгоритма является Gradient Boosting. Сущность алгоритма заключается в итеративной минимизации среднего квадрата ошибки прогнозов с использованием метода градиентного спуска.

Рассмотрим на примере использование описанных алгоритмов для решения задачи прогнозирования объемов продаж.

В 2013 г. компания «Волмарт» разместила в свободном доступе данные о еженедельных продажах 45-ти своих магазинов за период с 05. 02. 2010 г. по 01. 11. 2012 г. вместе с набором показателей, предположительно способных повлиять на уровень продаж. Сюда входят сведения об индексе потребительских цен, стоимости топлива на указанную дату, уровне безработицы, температуре воздуха, а также то, является ли неделя праздничной (например, предрождественской).

Очевидно, поиск зависимостей между указанными факторами является сложной статистической задачей, поскольку даже оценка самого факта влияния показателя на значения спроса может быть крайне трудоемкой и требовать построения комплексных корреляционных матриц.

В качестве основного инструмента для решения задачи используем язык программирования Python 3.8, широко применяемый в настоящее время в области искусственного интеллекта и машинного обучения. Целью исследования является создание модели, пригодной при прогнозировании объема продаж для любого из представленных магазинов.

Поскольку в настоящее время существует множество готовых библиотек в области машинного обучения, основными задачами, которые необходимо решить для построения корректной модели, является, прежде всего, подготовка подходящего набора данных, включающих достаточный набор показателей для каждого отдельно взятого случая, а также подбор основных параметров алгоритма.

В ходе исследования будут использованы данные о продажах 45-ти магазинов, в каждом из которых выделено до 99-ти товарных групп. Важно понимать, что влияние внешних факторов на объем продаж является не абсолютным, а относительным. Очевидно, что идентичные значения температуры воздуха, уровня безработицы и остальных показателей не гарантируют идентичные объемы продаж в разных магазинах. На это оказывает влияние, к примеру, расположение магазина — в центре города продажи будут выше, чем за его пределами. Сюда же можно отнести и размер магазина — более крупный объект будет иметь более высокие продажи.

В связи с этим наиболее корректным решением было бы создание отдельной модели для каждого магазина. Однако это решение потребовало бы подготовки и обучения 45-ти моделей для используемого набора данных, что, очевидно, является трудозатратным и к тому же не подлежит масштабированию.

Целью исследования в данном случае является поиск универсального решения, пригодного для использования с любым магазином. Соответственно, необходимо сопоставлять не только имеющиеся внешние факторы с объемом продаж, но и учитывать некоторые средние значения продаж отдельно взятого магазина за рассматриваемый период.

Рассчитаем среднее значение продаж каждого из магазинов в заданном интервале и проранжируем их от большего к меньшему. Полученный ранг будет использоваться как дополнительный входной параметр в процессе обучения. В результате должна быть получена математическая модель, использующая некоторые средние показатели продаж как базовые значения и определяющая отклонения от них на основе внешних факторов.

Дополнительно необходимо подготовить сами данные, заменяя все нечисловые значения числовыми. Например, вместо даты можно использовать порядковый номер недели, предпраздничные недели обозначим единицами в соответствующем столбце таблицы, остальные — нулями.

Итоговый набор факторов выглядит следующим образом (табл. 1).

Таблица 1. **Фрагмент набора факторов для анализа**

Магазин	Неделя	Температура воздуха, °C	Индекс стоимости топлива	ИПЦ	Уровень безработицы	Праздничная неделя	Ранг
1	1	5,7	2,572	211,0964	8,106	0	13
1	2	3,6	2,548	211,2422	8,106	1	13
1	3	4,4	2,514	211,2891	8,106	0	13
1	4	8,1	2,561	211,3196	8,106	0	13
1	5	7,9	2,625	211,3501	8,106	0	13

Примечание: источник [4].

Математические модели, лежащие в основе алгоритмов машинного обучения, как правило, оперируют числами в диапазоне [0, 1] или [–1, 1]. Приведение всех параметров к подобному формату возможно за счет деления единицы на заданное значение в том случае, когда значение выходит за пределы диапазона. Этот процесс называется нормализацией.

Важно отметить, что решение о включении того или иного показателя в набор данных зачастую является интуитивным. К примеру, зависимость между температурой воздуха и объемом продаж может быть достаточно слабой. Однако температура может существенным образом повлиять на то, какие именно товары приобретаются и таким образом перераспределять продажи между товарными группами, поэтому ее включение в список показателей является обоснованным.

Важное преимущество машинного обучения заключается именно в способности самостоятельного сопоставления входных параметров с выходными, и оценка их влияния на результат. Иначе говоря, если в процессе обучения модели будет выявлена слабая корреляция между температурой воздуха и объемом продаж, алгоритм учтет это, и при построении прогнозов температура также будет оказывать слабое влияние на результат. Это позволяет включать в модель дополнительные факторы, не беспокоясь о том, что в реальности они оказывают слабое воздействие на результат и их учет может сделать прогноз менее точным.

Однако не следует стремиться к максимизации числа факторов в надежде на то, что алгоритм самостоятельно определит самые важные параметры. Всегда необходимо учитывать соотношение объема данных с числом факторов, дабы быть уверенным, что данных достаточно для исключения вероятности выявления некорректных корреляций, отсутствующих в реальной жизни, но случайным образом прослеживающихся на отдельной выборке.

Используем следующий подход к рассматриваемой задаче. Разделим весь имеющийся набор данных на две части. Первая часть будет использована для обучения, вторая — для проверки результатов. Доли набора данных для обучения и тестирования составят соответственно 70 и 30 %. После обучения используем значения внешних факторов из тестовой выборки для расчета на их основе прогнозируемых значений и сравним полученные результаты с реальными показателями объемов продаж тестовой выборки. Оценку точности прогнозов проведем с помощью коэффициента детерминации R^2.

Построим прогноз для каждого из 45-ти магазинов, найдем суммарное значение этих прогнозов и сопоставим его с реальным значением суммарных продаж тестовой выборки.

Результат применения алгоритма Random Forest представлен на рис. 1.

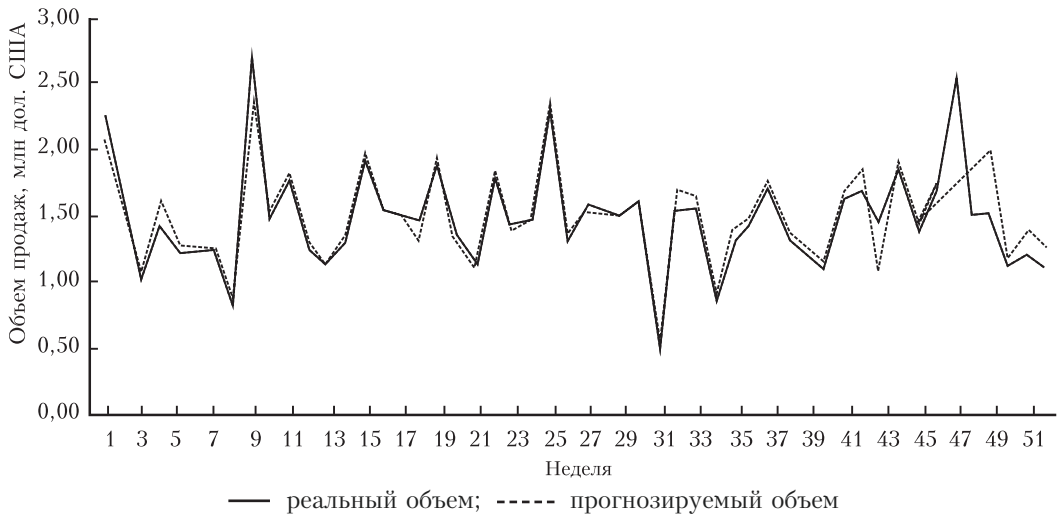

Рис. 1. Сравнение реальных и спрогнозированных объемов продаж при использовании алгоритма Random Forest

При использовании алгоритма Random Forest для оптимизации доступно широкое множество параметров, тем или иным образом влияющих на точность модели. Сложность применения алгоритма заключается в необходимости глубокого понимания специфики работы для оптимального подбора параметров модели. Рассмотрим ключевые из них.

Число деревьев — показатель, максимизация которого, как правило, приводит к повышению точности модели, одновременно увеличивая вычислительную сложность. Однако после определенного значения точность модели практически перестает повышаться, что делает нецелесообразным дальнейшее увеличение.

Для построения каждого разветвления в дереве просматривается определенное *число случайных признаков*. Это число также является регулируемым, по умолчанию используется значение квадратного корня из общего числа доступных признаков. Поскольку в рассматриваемом случае доступно всего 7 признаков, используем значение по умолчанию.

Максимальная глубина деревьев — так же, как и число деревьев, дает более высокую точность при увеличении до определенного предела. Однако существует сценарий, при котором рекомендуется использовать большое количество неглубоких деревьев, в задачах с высоким числом шумовых объектов (выбросов).

В табл. 2 представлены результаты прогнозов с использованием различных значений параметров. Наилучшее значение R^2 в 0,835, означающее, что прогноз на 83,5 % совпал с реальным значением, было достигнуто при оценке с использованием четырехсот деревьев, максимальной глубиной дерева 48. Дальнейшее увеличение числа деревьев и глубины не приводит к существенному повышению точности [2, с. *340*].

Таблица 2. **Результаты применения алгоритма Random Forest**

Число деревьев	Максимальная глубина деревьев	R^2
200	24	0,802
300	32	0,824
400	48	0,835

Алгоритм Gradient Boosting, в основе которого также лежат деревья решений, оперирует тем же набором параметров, что и Random Forest. Однако подбор оптимальных значений в этом случае следует вести с большей осторожностью, так как алгоритм в большей степени подвержен проблеме переобучения

(Overfitting). Суть этого явления состоит в том, что модель слишком хорошо объясняет примеры из обучающей выборки, но гораздо хуже работает на примерах, не участвовавших в обучении. Чтобы избежать проблемы, следует более осторожно увеличивать число деревьев и их максимальную глубину (рис. 2).

Рис. 2. Сравнение реальных и спрогнозированных объемов продаж при использовании алгоритма Gradient Boosting

В целом, алгоритм Gradient Boosting в рассмотренной задаче показывает себя несколько хуже, однако также дает результаты с точностью выше 70 % (табл. 3).

Таблица 3. **Результаты применения алгоритма Gradient Boosting**

Число деревьев	Максимальная глубина деревьев	R^2
100	24	0,709
150	32	0,722
200	48	0,745

Заключение. Таким образом, в рамках представленной статьи была рассмотрена сущность алгоритмов машинного обучения, проанализированы их преимущества и недостатки в сравнении с традиционными методиками прогнозирования и описаны возможности их применения для построения прогнозов объемов продаж.

С помощью алгоритмов Gradient Boosting и Random Forest были приведены примеры подготовки и нормализации данных для обучения модели, выделения обучающей и тестовой выборки, оценки результатов применения алгоритмов. Следует отметить, что рассмотренная последовательность шагов является универсальной для практики применения алгоритмов машинного обучения и не зависит от конкретного набора данных, используемых для анализа.

Приведенные примеры использования алгоритмов для построения прогнозов объемов продаж розничной торговой сети «Волмарт» позволяют утверждать, что точность полученных прогнозов превышает 70 % на рассмотренном наборе данных. Надо отметить высокую востребованность подобных подходов и возрастающий интерес к ним в области решения широкого спектра экономико-математических задач. Важно понимать, что приведенные примеры не являются исчерпывающими и в реальной практике зачастую используют комбинации различных подходов, а число доступных алгоритмов крайне велико. Кроме того, само понятие «машинное обучение» является крайне многогранным и включает в себя множество современных областей, например, теорию нейронных сетей.

В настоящее время применение алгоритмов машинного обучения еще не получило широкого распространения, а сами алгоритмы используются для решения специфического круга задач, как правило, в крупных компаниях. В большинстве случаев они применяются совместно с традиционными методами прогнозирования. Тем не менее стремительно возрастающие вычислительные мощности обусловливают возможность более эффективного применения машинного обучения в практике управления цепями поставок, а значит, можно прогнозировать рост востребованности этой области в обозримом будущем.

Литература и электронные публикации в Интернете

1. Demand Forecasting Methods: Using Machine Learning and Predictive Analytics to See the Future of Sales [Electronic resource] // Altexsoft Software R&D Engineering. — Mode of access: https://www.altexsoft.com/blog/demand-forecasting-methods-using-machine-learning/. — Date of access: 29.03.2020.

2. *Флах, П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — М. : ДМК Пресс, 2019. — 403 с.

Flakh, P. Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh [Maching Learning. The Science and Art of building algorithms that extract knowledge from data] / P. Flakh. — M. : DMK Press, 2019. — 403 p.

3. *Вьюгин, В. В.* Математические основы машинного обучения и прогнозирования / В. В. Вьюгин. — М. : Издательство МЦНМО, 2014. — 303 с.

V'yugin, V. V. Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya [Mathematical Foundations of Maching Learning and Forecasting] / V. V. V'yugin. — M. : Izdatel'stvo MTsNMO, 2014. — 303 p.

4. Walmart Recruiting — Store Sales Forecasting [Electronic resource]. — Mode of access: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/. — Date of access: 20.03.2020.

SIARHEI HAURYLAU

ADVANTAGES OF MATHEMATICAL FORECASTING TOOLS

Author affiliation. *Siarhei HAURYLAU* (gavrilovsv_95@mail.ru), *Belarus State Economic University (Minsk, Belarus).*

Abstract. The article analyzes operating principles of a number of machine learning algorithms, describes the scripts of their application in solving forecasting tasks, based on the example of retailing sphere. The main result of the research is confirmation of the opportunity to implement machine learning tools in the forecasting practice, as well as the proof of the high effectiveness of these algorithms.

Keywords: machine learning; sales forecasting; inventory management.

UDC 338.984

Статья поступила
в редакцию 31. 08. 2020 г.