Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2022, №6, Том 14 / 2022, No 6, Vol 14 https://esj.today/issue-6-2022.html

URL статьи: https://esj.today/PDF/52ECVN622.pdf

Ссылка для цитирования этой статьи:
Атрохова, А. Н. Методы машинного обучения для прогнозирования продаж в пункте выдачи заказов Wildberries / А. Н. Атрохова // Вестник евразийской науки. — 2022. — Т. 14. — № 6. — URL: https://esj.today/PDF/52ECVN622.pdf

For citation:
Anastasia A.N. Machine learning methods for predicting sales at the Wildberries order pick-up point. *The Eurasian Scientific Journal*. 2022; 14(6): 52ECVN622. Available at: https://esj.today/PDF/52ECVN622.pdf. (In Russ., abstract in Eng.).

Атрохова Анастасия Николаевна
ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, Россия
2 курс магистратуры
E-mail: sserenityy@mail.ru
РИНЦ: https://elibrary.ru/author_profile.asp?id=1032653

Методы машинного обучения для прогнозирования продаж в пункте выдачи заказов Wildberries

Аннотация. Большая часть современной экономики основана на информации, поэтому различные виды электронной коммерции становятся очень востребованными и постепенно вытесняют физическую коммерцию, а торговые интернет-площадки с развитой сетью пунктов выдачи заказов становятся более распространенными.

Российские маркетплейсы предлагают на правах партнера открыть пункт выдачи заказов. Предпринимателю, желающему использовать эту опцию и создать такой бизнес, необходимо использовать предиктивную аналитику, поскольку большинство крупных маркетплейсов, таких как Wildberries, переводят партнеров на процент от оборота, уходя от фиксированной ставки за выдачу. Таким образом, выручка собственника пункта выдачи заказов зависит от объема продаж на конкретной точке, ввиду чего представляется актуальным необходимость прогнозировать будущий объем продаж.

Объектом исследования выступает методика прогнозирования продаж в сфере ритейла с помощью алгоритмов машинного обучения, а предметом — автоматизированные инструменты прогнозирования объема продаж в пунктах выдачи заказов Wildberries.

Автором представлены метрики качества моделей по протестированным 6 методам машинного обучения: Linear Regression, Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Neural Network Regression, Poisson Regression.

Результатом настоящего исследования является самостоятельно разработанная автором предиктивная модель объема продаж пункта выдачи заказов Wildberries. Практическая значимость работы обусловлена возможностью применения данной модели для прогнозирования выручки предпринимателя-собственника пункта выдачи заказов в целях дальнейшего сопоставления предиктивных данных со статьей расходов и анализа прибыльности действующего пункта выдачи заказов.

Главное отличие и новизна данной работы состоит в том, что ранее ни одно из исследований не было рассмотрено с позиции интересов собственника пункта выдачи заказов.

Ключевые слова: прогнозирование продаж; предиктивная аналитика; методы машинного обучения; регрессия; обработка данных; анализ данных; пункт выдачи заказов; ритейл

Введение

Предпочтения конечных получателей в сфере e-commerce движутся в сторону самовывоза и с каждым годом его доля растет. Сейчас в среднем по рынку соотношение курьерской доставки и самостоятельного забора по всем категориям товаров составляет 40 % на 60 %. По прогнозам экспертов, доля самовывоза продолжит увеличиваться[1].

Рынок пунктов выдачи заказов (далее — ПВЗ) как узкая ниша сферы ритейла интересен тем, что он объединяет преимущества электронной коммерции в части огромного ассортимента площадок, выгодных цен, доступности, и традиционной торговли с примеркой, простым возвратом, частичным выкупом заказа.

В отечественной литературе отмечается ряд авторов, исследующих методы машинного обучения для прогнозирования продаж в ритейле: Виноградов В.И. [1], Мокшин В.В. [2], Сологуб Г.Б. [3], Каипов И.К. [4], Сердинская Ю.А. [5], Инюцина В.С. [6] и др. Однако ни одна из работ не рассмотрена с позиции интересов собственника пункта выдачи заказов в части максимизации его прибыли в результате ведения малого бизнеса.

Задачи исследования:

1. Раскрыть категориальный аппарат исследования.
2. Исследовать возможности применения алгоритмов машинного обучения в пунктах выдачи заказов Wildberries.
3. Выбрать атрибуты и целевую переменную для построения предиктивной модели объема продаж в пункте выдачи заказов Wildberries.
4. На основе больших данных построить предиктивную модель объема продаж в пункте выдачи заказов Wildberries.
5. Оценить предложенную модель с точки зрения ее адекватности на основе анализа метрик качества модели.

Digital-трансформация бизнес-процессов охватывает все сферы экономики, в том числе e-commerce. Цифровая трансформация бизнеса осуществляется с целью создания улучшенных и стабильных бизнес-моделей для организаций, которые могут продуктивно работать в условиях современной цифровой промышленной политики и адаптироваться к изменениям [7].

Главной задачей ритейл-бизнеса в высококонкурентной среде является необходимость привлечения и удержания потребителей за счет удовлетворения их потребностей, а также сокращение расходов и максимизации прибыли.

Различные решения на основе технологий искусственного интеллекта помогают улучшить систему финансового и бизнес-планирования.

Главной проблемой, препятствующей внедрению новых технологий, является высокая стоимость их обслуживания [8].

[1] Перегрет ли рынок ПВЗ, трудности при открытии точек, что будет дальше. Интервью с генеральным директором Hermes Russia Алексеем Шулевым. URL: https://oborot.ru/articles/pvz-interview-15-i153835.html?ysclid=l88kwtrtlq602881940 (дата обращения 16.10.2022).

Прогнозную аналитику в розничной торговле обычно могут позволить себе компании с миллиардными оборотами, такие как, например, Wildberries. Это обусловлено тем, что при работе с прогнозными моделями требуются знания в области больших данных, искусственного интеллекта и машинного обучения. Процесс подразумевает не одну программу, а целый набор методологий работы с данными[2].

Открытие пункта выдачи заказов в качестве партнера одного из маркетплейсов относится к малому бизнесу. Это достаточно узкая ниша сферы ритейла, но для эффективной деятельности такому бизнесу, как и любому другому, требуется иметь доступ к прогнозной аналитике. Собственнику пункта выдачи заказов также необходимы финансово доступные методы и инструменты планирования и прогнозирования объема продаж и выручки, поскольку такой бизнес под влиянием внутренних и внешних экономических факторов может перестать быть прибыльным. В таком случае представляется важным вовремя принять решение о целесообразности сохранения бизнеса во избежание несения больших убытков.

Методы исследования: моделирование, анализ, сравнение, дедукция и обобщение.

Машинное обучение

Среди используемых в международном ритейле технологий искусственного интеллекта лидерство принадлежит машинному обучению. По данным опроса Capgemini Research Institute[3], машинное обучение используют 67 % ритейлеров, которые в той или иной степени внедрили искусственный интеллект в свои бизнес-процессы. В тройку технологий-лидеров также входят обработка естественного языка — 14 % и компьютерное зрение — 13 %. В российском ритейле машинное обучение также считается одной из основных применяемых технологий искусственного интеллекта[4].

Машинное обучение (ML) — это использование математических моделей данных, которые помогают компьютеру обучаться без непосредственных инструкций. При машинном обучении с помощью алгоритмов выявляются закономерности в данных, на основе которых создается прогнозная модель. Чем больше данных обрабатывает такая модель и чем дольше она используется, тем точнее становятся результаты[5].

Процесс машинного обучения начинается со сбора набора данных, показывающего связь между целевыми и другими признаками для большого количества сущностей одной категории. Затем значения признаков вводятся в модель, и вычисляется теоретическое значение цели по отношению к другим признакам. Важно, чтобы полученные значения прогноза целевой переменной были близки к ее истинным значениям, так как от этого напрямую зависит эффективность алгоритма [9]. Разрыв между теоретическим значением и измеренным значением рассчитывается как «ошибка», и модель корректирует свои параметры, чтобы

[2] Predictive Analytics in Retail & E-commerce. January 7, 2021. URL: https://indatalabs.com/blog/predictive-analytics-in-retail-and-e-commerce (date of application 16.10.2022).

[3] Building the Retail Superstar: How unleashing AI across functions offers a multi-billion dollar opportunity, Capgemini Research Institute. // URL: https://www.capgemini.com/wp-content/uploads/2018/12/Report-%E2%80%93-Building-the-Retail-Superstar-Digital1.pdf (date of application 17.10.2022).

[4] Исследование РАЭК / НИУ ВШЭ при поддержке Microsoft «Искусственный интеллект в ритейле: практика российского бизнеса». URL: https://raec.ru/activity/analytics/11479/ (дата обращения 17.10.2022).

[5] Что такое машинное обучение и как оно работает. Официальный сайт облачной платформы Azure. URL: https://azure.microsoft.com/ru-ru/resources/cloud-computing-dictionary/what-is-machine-learning-platform/ (дата обращения 17.10.2022).

минимизировать ошибку [10]. Такой тип машинного обучения называется «обучение с учителем».

В машинном обучении используются многие типы данных, но их можно условно разделить на количественные или качественные данные.

Набор данных, качество и продолжительность периода, за который они собраны, определяют, насколько точная модель получится в итоге.

Для того чтобы нейросети правильно строили модели, нужно собирать достоверные данные, тщательно очищать их от постороннего шума и подготавливать для машинного обучения. Этап подготовки называют предпроцессингом — информацию переводят в формат, подходящий для обучения алгоритма[6].

Прогнозирование объема продаж в пункте выдачи заказов Wildberries

Для целей настоящего исследования из Личного кабинета собственника WB Point был собран обезличенный набор данных (далее — датасет) по пункту выдачи заказов Wildberries, расположенному по адресу г. Москва, ул. Толбухина 13к1, за период 01.09.2021–18.11.2021 объемом 3329 наблюдений[7].

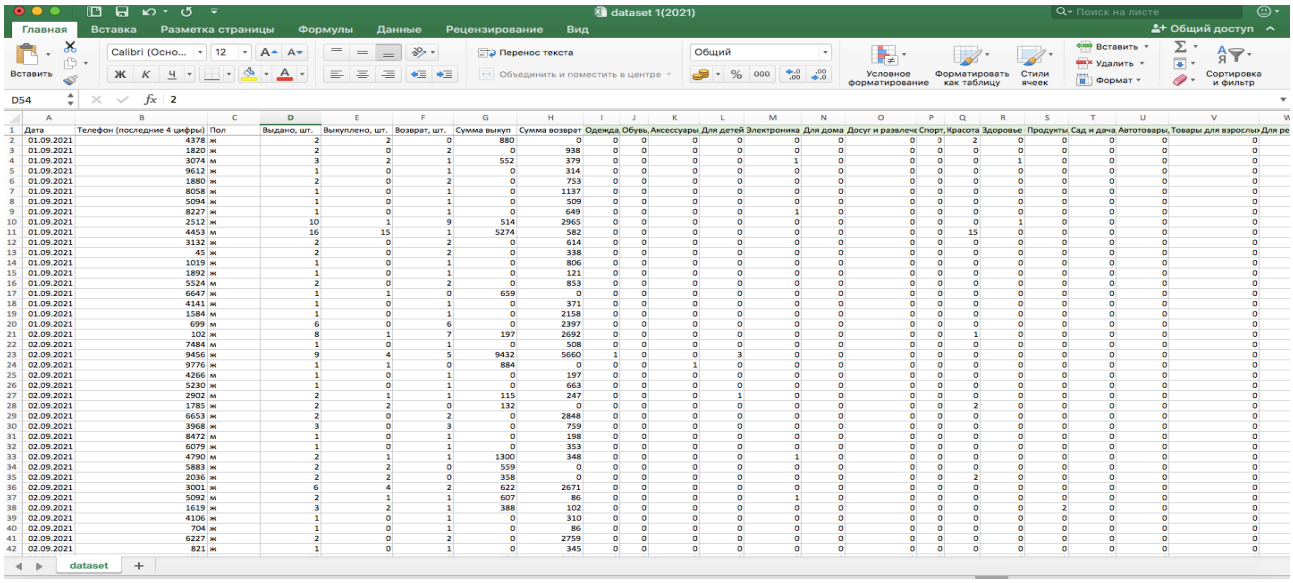
Набор данных содержит 64 атрибута (рис. 1):

- *10 общих переменных:* дата; телефон (последние 4 цифры); пол; выдано, шт.; выкуплено, шт.; возврат, шт.; сумма выкупа, руб.; сумма возврата, руб.; количество клиентов в ПВЗ за день; количество выданных товаров в ПВЗ за день.
- *18 переменных по категориям выкупленных товаров в штуках*: одежда, выкуп. шт.; обувь, выкуп. шт.; аксессуары, выкуп. шт.; для детей, выкуп. шт.; электроника, выкуп. шт.; для дома, выкуп. шт.; досуг и развлечения, выкуп. шт.; спорт, выкуп. шт.; красота, выкуп. шт.; здоровье, выкуп. шт.; продукты, выкуп. шт.; сад и дача, выкуп. шт.; автотовары, выкуп. шт.; товары для взрослых, выкуп. шт.; для ремонта, выкуп. шт.; книги, выкуп. шт.; канцтовары, выкуп. шт.; зоотовары, выкуп. шт.
- *18 переменных по категориям выкупленных товаров в рублях:* одежда, выкуп. руб.; обувь, выкуп. руб.; аксессуары, выкуп. руб.; для детей, выкуп. руб.; электроника, выкуп. руб.; для дома, выкуп. руб.; досуг и развлечения, выкуп. руб.; спорт, выкуп. руб.; красота, выкуп. руб.; здоровье, выкуп. руб.; продукты, выкуп. руб.; сад и дача, выкуп. руб.; автотовары, выкуп. руб.; товары для взрослых, выкуп. руб.; для ремонта, выкуп. руб.; книги, выкуп. руб.; канцтовары, выкуп. руб.; зоотовары, выкуп. руб.
- *18 переменных по категориям возвратных товаров в штуках*: одежда, возврат шт.; обувь, возврат шт.; аксессуары, возврат шт.; для детей, возврат шт.; электроника, возврат шт.; для дома, возврат шт.; досуг и развлечения, возврат шт.; спорт, возврат шт.; красота, возврат шт.; здоровье, возврат шт.; продукты, возврат шт.; сад и дача, возврат шт.; автотовары, возврат шт.; товары для взрослых,

[6] Как Machine Learning повышает продажи. Журнал VK Cloud об IT-бизнесе, технологиях и цифровой трансформации. URL: https://mcs.mail.ru/blog/kak-machine-learning-povyshaet-prodazhi (дата обращения 17.10.2022).

[7] Личный кабинет для собственников и сотрудников пунктов выдачи заказов Wildberries. URL: https://point.wb.ru/ (дата обращения 17.10.2022).

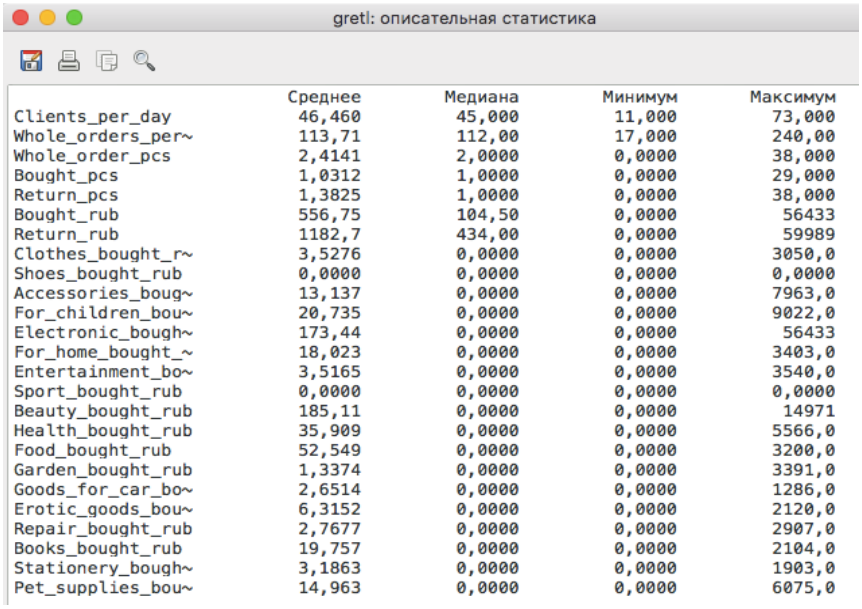
возврат шт.; для ремонта, возврат шт.; книги, возврат шт.; канцтовары, возврат шт.; зоотовары, возврат шт.



Дата	Телефон (последние 4 цифры)	Пол	Выдано, шт.	Выкуплено, шт.	Возврат, шт.	Сумма выкуп	Сумма возврат	Одежда	Обувь	Аксессуары	Для детей	Электроника	Для дома	Досуг и развлече	Спорт	Красота	Здоровье	Продукты	Сад и дача	Автотовары	Товары для взрослых	Для ре
01.09.2021	4378	ж	2	2	0	880	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
01.09.2021	1820	ж	2	0	2	0	938	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	3074	м	3	2	1	552	379	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
01.09.2021	9612	ж	1	0	1	0	314	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	1880	ж	2	0	2	0	753	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	8058	ж	1	0	1	0	1137	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	5094	ж	1	0	1	0	509	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	8227	ж	1	0	1	0	649	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
01.09.2021	2512	ж	10	1	9	514	2965	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
01.09.2021	4453	м	16	15	1	5274	582	0	0	0	0	0	0	0	0	15	0	0	0	0	0	
01.09.2021	3132	ж	2	0	2	0	614	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	45	ж	2	0	2	0	338	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	1019	ж	1	0	1	0	806	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	1892	ж	1	0	1	0	121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	5524	м	2	0	2	0	853	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	6647	ж	1	1	0	659	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	4141	ж	1	0	1	0	371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	1584	м	1	0	1	0	2158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
01.09.2021	691	м	6	0	6	0	2397	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	102	ж	8	1	7	197	2692	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
02.09.2021	7484	м	1	0	1	0	508	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	9456	ж	9	4	5	9432	5660	1	0	0	3	0	0	0	0	0	0	0	0	0	0	
02.09.2021	9776	ж	1	1	0	884	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	4266	м	1	0	1	0	197	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	5230	ж	1	0	1	0	663	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	2902	ж	2	1	1	115	247	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
02.09.2021	1783	ж	2	2	0	132	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
02.09.2021	6653	ж	2	0	2	0	2848	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	3968	ж	3	0	3	0	759	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	8472	м	1	0	1	0	198	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	6079	ж	1	0	1	0	353	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	4790	м	2	1	1	1300	348	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
02.09.2021	5883	ж	2	2	0	559	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	2036	ж	2	2	0	358	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
02.09.2021	3001	ж	6	4	2	622	2671	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	5092	м	2	1	1	607	86	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
02.09.2021	1619	ж	3	2	1	388	102	0	0	0	0	0	0	0	0	0	0	2	0	0	0	
02.09.2021	4106	ж	1	0	1	0	310	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	704	ж	1	0	1	0	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	6227	ж	2	0	2	0	2759	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
02.09.2021	821	ж	1	0	1	0	345	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Рисунок 1. *Набор данных за период 01.09.2021–18.11.2021 объемом 3329 наблюдений. Составлено автором самостоятельно на основе данных из Личного кабинета собственника WB Point*

Используя программное обеспечение Gretl, приведем описательную статистику по набору данных за период 01.09.2021–18.11.2021 объемом 3329 наблюдений (рис. 2). Так, за исследуемый период пункт выдачи заказов Wildberries по адресу г. Москва, ул. Толбухина 13к1 каждый день в среднем посещает 46 человек, среднее количество выданных заказов составляет 114 товаров в день. Максимальное количество выкупленных товаров за анализируемый период — 29, возвращенных — 38. Средняя сумма выкупа по каждому заказу равна 556,75 рублей, средняя сумма возврата — 1182,7 руб.



gretl: описательная статистика

	Среднее	Медиана	Минимум	Максимум
Clients_per_day	46,460	45,000	11,000	73,000
Whole_orders_per~	113,71	112,00	17,000	240,00
Whole_order_pcs	2,4141	2,0000	0,0000	38,000
Bought_pcs	1,0312	1,0000	0,0000	29,000
Return_pcs	1,3825	1,0000	0,0000	38,000
Bought_rub	556,75	104,50	0,0000	56433
Return_rub	1182,7	434,00	0,0000	59989
Clothes_bought_r~	3,5276	0,0000	0,0000	3050,0
Shoes_bought_rub	0,0000	0,0000	0,0000	0,0000
Accessories_boug~	13,137	0,0000	0,0000	7963,0
For_children_bou~	20,735	0,0000	0,0000	9022,0
Electronic_bough~	173,44	0,0000	0,0000	56433
For_home_bought_~	18,023	0,0000	0,0000	3403,0
Entertainment_bo~	3,5165	0,0000	0,0000	3540,0
Sport_bought_rub	0,0000	0,0000	0,0000	0,0000
Beauty_bought_rub	185,11	0,0000	0,0000	14971
Health_bought_rub	35,909	0,0000	0,0000	5566,0
Food_bought_rub	52,549	0,0000	0,0000	3200,0
Garden_bought_rub	1,3374	0,0000	0,0000	3391,0
Goods_for_car_bo~	2,6514	0,0000	0,0000	1286,0
Erotic_goods_bou~	6,3152	0,0000	0,0000	2120,0
Repair_bought_rub	2,7677	0,0000	0,0000	2907,0
Books_bought_rub	19,757	0,0000	0,0000	2104,0
Stationery_bough~	3,1863	0,0000	0,0000	1903,0
Pet_supplies_bou~	14,963	0,0000	0,0000	6075,0

Рисунок 2. *Gretl: описательная статистика по набору данных за период 01.09.2021–18.11.2021 объемом 3329 наблюдений. Составлено автором самостоятельно на основе данных из Личного кабинета собственника WB Point*

52ECVN622
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Максимальная сумма выкупленных товаров по категории «Одежда» — 3050 рублей, по категории «Обувь» — 0 рублей, по категории «Аксессуары» — 7963 рубля, по категории «Для детей» — 9022 рубля, по категории «Электроника» — 56433 рубля, по категории «Для дома» — 3403 рубля, по категории «Развлечения» — 3540 рублей, по категории «Спорт» — 0 рублей, по категории «Красота» — 14971 рубль, по категории «Здоровье» — 5566 рублей, по категории «Продукты» — 3200 рублей, по категории «Сад и дача» — 3391 рубль, по категории «Автотовары» — 1286 рублей, по категории «Товары для взрослых» — 2120 рублей, по категории «Для ремонта» — 2907 рублей, «Книги» — 2104 рубля, по категории «Канцтовары» — 1903 рубля и по категории «Зоотовары» — 6075 рублей.

Для прогнозирования будущего объема продаж определена целевая переменная «Сумма выкупа, руб.». В данной работе для построения моделей машинного обучения «с учителем» используется студия машинного обучения Microsoft Azure Machine Learning Studio[8] (рис. 2).

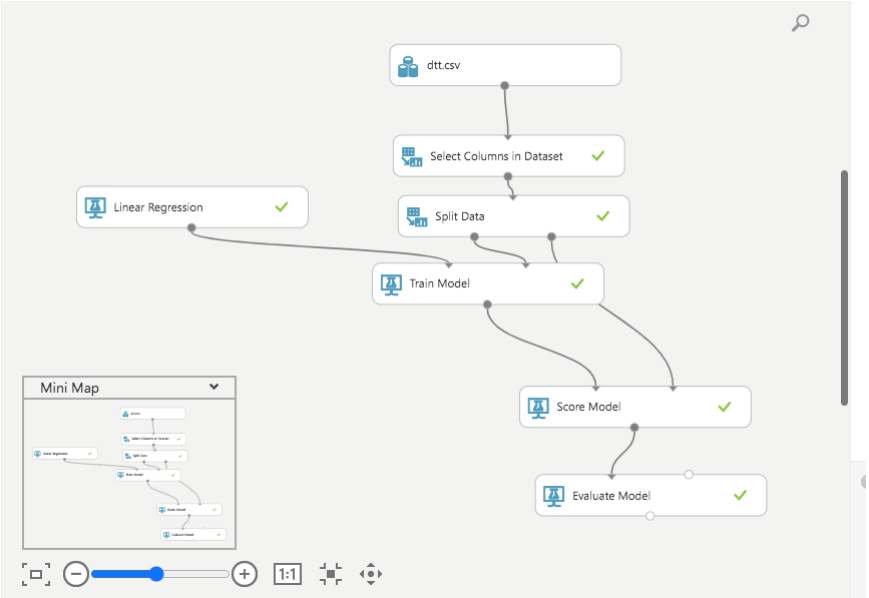


Рисунок 3. *Предиктивная модель объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод линейной регрессии). Составлено автором самостоятельно*

Эксперимент 1

В целях построения предиктивной модели был выбран такой метод машинного обучения, как линейная регрессия. Поскольку исходный набор данных имеет 64 переменных, автором было построено 7 моделей для выявления наиболее значимых признаков, влияющих на качество прогнозирования:

- модель 1 со всеми собранными признаками (64 атрибута);
- модель 2 с общими переменными и переменными по категориям выкупленных товаров в рублях (28 атрибутов);
- модель 3 с общими переменными, переменными по категориям выкупленных товаров в рублях и переменными по категориям возвратных товаров в штуках (46 атрибутов);

[8] Microsoft Azure Machine Learning Studio. URL: https://studio.azureml.net/ (date of application 16.10.2022).

- модель 4 с общими переменными, переменными по категориям выкупленных товаров в рублях и переменными по категориям выкупленных товаров в штуках (46 атрибутов);
- модель 5 с общими переменными, переменными по категориям выкупленных товаров в штуках и переменными по категориям возвратных товаров в штуках (46 атрибутов);
- модель 6 с общими переменными и переменными по категориям возвратных товаров в штуках (28 атрибутов);
- модель 7 с общими переменными без категоризации (10 атрибутов).

Результаты моделирования представлены в таблице 1.

Таблица 1

Метрики качества полученных предиктивных моделей объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод линейной регрессии), обученных на данных за период 01.09.2021–18.11.2021, выборка поделена на обучающую и тестовую в соотношении 70 на 30 (Split Data: Fraction of rows in the first output dataset = 0,7)

	Модель 1	Модель 2	Модель 3	Модель 4	Модель 5	Модель 6	Модель 7
Mean Absolute Error	2.141435	1.729623	1.77745	2.071635	557.78552	581.16298	573.704872
Root Mean Squared Error	34.13549	34.147392	34.1287	34.17709	2663.5748	2711.16521	2710.28298
Relative Absolute Error	0.002441	0.001972	0.002026	0.002361	0.635796	0.662443	0.653942
Relative Squared Error	0.000149	0.000149	0.000149	0.000149	0.907103	0.939807	0.939196
Coefficient of Determination	0.999851	0.999851	0.999851	0.999851	0.092897	0.060193	0.060804

Составлено автором самостоятельно на основе данных Microsoft Azure Machine Learning Studio

Полученные метрики качества позволяют судить об адекватности построенных моделей. Так, чем ближе коэффициент детерминации к единице, тем лучше регрессия аппроксимирует данные и тем точнее предиктивная модель. Самый высокий коэффициент детерминации наблюдается у моделей 1–4: R^2 = 0,999851. У всех четырех моделей одинаковая относительная стандартная ошибка: RSE = 0,000149. Наименьшие средняя абсолютная ошибка (MAE) и относительная абсолютная ошибка (RAE) отмечаются у модели 2, при этом среднеквадратичная ошибка данной модели несколько выше, чем у моделей 1 и 3: RMSE = 34,14739. По совокупности полученных метрик качества наилучшей для прогнозирования объема продаж в пунктах выдачи заказов Wildberries является модель 2 с общими переменными и переменными по категориям выкупленных товаров в рублях (28 атрибутов).

Для построения предиктивной модели на базе алгоритмов машинного обучения можно также использовать другие методы, такие как: Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Neural Network Regression. Poisson Regression.

Сравним метрики качества полученной предиктивной модели 2, построенной по методу линейной регрессии, с метриками качества прогнозных моделей, в основе которых лежит применение каждого из вышеперечисленных пяти методов. Набор атрибутов для всех шести моделей идентичен: для их построения использованы общие переменные и переменные по категориям выкупленных товаров в рублях (28 атрибутов).

Метрики качества шести регрессионных моделей, каждая из которых на вход получает одинаковые данные и различается только методом машинного обучения, отображены в таблице 2.

Таблица 2

Метрики качества полученных предиктивных моделей объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод линейной регрессии, метод байесовской линейной регрессии, метод дерева решений, метод случайного леса, метод нейронных сетей, метод пуассоновской регрессии), обученных на данных за период 01.09.2021–18.11.2021, выборка поделена на обучающую и тестовую в соотношении 70 на 30 (Split Data: Fraction of rows in the first output dataset = 0,7)

	Linear Regression	Bayesian Linear Regression	Decision Tree Regression	Decision Forest Regression	Neural Network Regression	Poisson Regression
Mean Absolute Error	1.729623	1.71971	259.593206	299.17199	702.373063	729.640044
Root Mean Squared Error	34.147392	34.14327	2153.077596	2323.314041	2839.332061	2731.461301
Relative Absolute Error	0.001972	0.00196	0.295899	0.341669	0.800605	0.831685
Relative Squared Error	0.000149	0.000149	0.592715	0.692109	1.030764	0.953931
Coefficient of Determination	0.999851	0.999851	0.407285	0.307891	-0.030764	0.046069

Составлено автором самостоятельно на основе данных Microsoft Azure Machine Learning Studio

Результаты оценки метрик качества из таблицы 2 позволяют сделать вывод о том, что наилучшими методами для прогнозирования объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения являются линейная регрессия и байесовская линейная регрессия, так как коэффициент детерминации близок к единице, что говорит о высокой точности моделей. При этом прочие метрики качества так же говорят об адекватности этих моделей. Выбирая один из двух методов, следует отдать предпочтение байесовской линейной регрессии, так как на выходе автор получил наилучшие параметры по каждой метрике: самая низкая средняя абсолютная ошибка MAE = 1,71971, самая низкая средняя абсолютная ошибка, самая низкая среднеквадратичная ошибка RMSE = 34,14327 и самая низкая относительная абсолютная ошибка RAE = 0,00196.

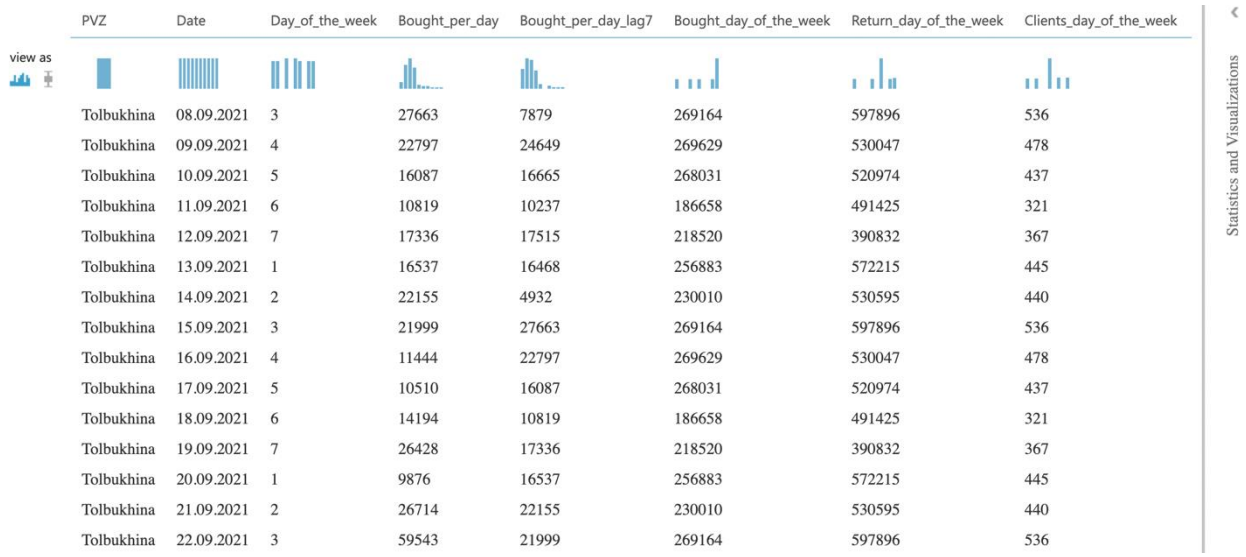
Также экспериментальным путем было выявлено, что категориальные данные никак не влияют на прогнозирование целевой переменной, поэтому в целях дальнейшего исследования их можно исключить из модели.

Стоит отметить, что построенная модель характеризуется высокой точностью в связи с тем, что она не содержит лаговых признаков, ввиду чего ее можно использовать только в условиях реального времени, подставляя любые интересующие значения и получая на выходе значение целевой переменной. Для построения предиктивной модели необходимо ввести в модель лаговые значения.

Эксперимент 2

В исходный датасет была добавлена лаговая переменная Bought_per_day_lag7, которая характеризует сумму выкупа товаров за 7 дней до соответствующей даты. Например, в наборе данных для 8 сентября 2021 года Bought_per_day_lag7 показывает сумму выкупленных товаров 1 сентября 2021 года, для 9 сентября 2021 года — сумму выкупленных товаров 2 сентября 2021 года и т. д. (рис. 4). Поскольку исходный набор данных в 3329 наблюдений содержал значения с 1 сентября 2021 года, то по первой неделе нет лаговых значений (так как нет данных

за период 25.08.2021–01.09.2021), поэтому новый датасет с лаговой переменной Bought_per_day_lag7 содержит наблюдения за период 08.09.2021–18.11.2021 гг.



PVZ	Date	Day_of_the_week	Bought_per_day	Bought_per_day_lag7	Bought_day_of_the_week	Return_day_of_the_week	Clients_day_of_the_week
Tolbukhina	08.09.2021	3	27663	7879	269164	597896	536
Tolbukhina	09.09.2021	4	22797	24649	269629	530047	478
Tolbukhina	10.09.2021	5	16087	16665	268031	520974	437
Tolbukhina	11.09.2021	6	10819	10237	186658	491425	321
Tolbukhina	12.09.2021	7	17336	17515	218520	390832	367
Tolbukhina	13.09.2021	1	16537	16468	256883	572215	445
Tolbukhina	14.09.2021	2	22155	4932	230010	530595	440
Tolbukhina	15.09.2021	3	21999	27663	269164	597896	536
Tolbukhina	16.09.2021	4	11444	22797	269629	530047	478
Tolbukhina	17.09.2021	5	10510	16087	268031	520974	437
Tolbukhina	18.09.2021	6	14194	10819	186658	491425	321
Tolbukhina	19.09.2021	7	26428	17336	218520	390832	367
Tolbukhina	20.09.2021	1	9876	16537	256883	572215	445
Tolbukhina	21.09.2021	2	26714	22155	230010	530595	440
Tolbukhina	22.09.2021	3	59543	21999	269164	597896	536

Рисунок 4. *Фрагмент датасета с лаговой переменной Bought_per_day_lag7 (Составлено автором самостоятельно)*

Также для эксперимента были добавлены такие признаки, как Day_of_the_week (день недели), Bought_day_of_the_week (количество выкупленных товаров в конкретный день недели), Return_day_of_the_week (количество возвращенных товаров в конкретный день недели), Clients_day_of_the_week (количество клиентов в конкретный день недели) и Orders_day_of_the_week (количество выданных заказов в конкретный день недели). Переменные «…_day_of_the_week» были рассчитаны путем суммирования количества товаров/заказов/клиентов, отфильтрованных по дням недели (отдельно — по понедельникам, вторникам, средам и т.д.), поэтому каждый из этих признаков характеризуется ровно 7 уникальными значениями. Кроме того, как видно из рис. 4, даты взяты агрегированно, то есть на каждую дату отводится лишь одна строка, поэтому количество наблюдений в данном датасете сокращено с 3329 до 72.

За целевую переменную так же, как и в предыдущем эксперименте, принята сумма выкупленных товаров за день «Bought_per_day». Метод линейной регрессии и настройка блока Split Data: Fraction of rows in the first output dataset = 0,7 показали следующие результаты (табл. 3).

Таблица 3

Метрики качества предиктивной модели объема продаж в пунктах выдачи заказов Wildberries с участием лаговой переменной. Датасет содержит 10 признаков: Date, Day_of_the_week, Bought_per_day (целевая переменная), Bought_per_day_lag7 (лаговая переменная), Clients_per_day, Whole_orders_per_day, Bought_day_of_the_week, Return_day_of_the_week, Clients_day_of_the_week, Orders_day_of_the_week

Mean Absolute Error (средняя абсолютная ошибка)	11550.047834
Root Mean Squared Error (среднеквадратичная ошибка)	14258.461864
Relative Absolute Error (относительная абсолютная ошибка)	1.230014
Relative Squared Error (относительная стандартная ошибка)	0.885775
Coefficient of Determination	0.114225

Составлено автором самостоятельно

52ECVN622
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Метрики качества неутешительны, поскольку они свидетельствуют о несостоятельности модели: R2 = 0,114. Предположительно, причиной такого результата является малый объем обучающей выборки, поэтому в дальнейших экспериментах принято решение использовать первичные неагрегированные данные в разрезе каждого дня, которые дополнены лаговой переменной.

Эксперимент 3

Для эксперимента № 3 взят датасет объемом 3119 наблюдений, дополненный двумя лаговыми переменными, со следующими 10 признаками: Date, Day_of_the_week, Gender (пол клиента), Whole_order_pcs, Bought_rub, Bought_pcs, Bought_per_day (целевая переменная), Bought_per_day_lag7 (лаговая переменная), Clients_per_day_lag7 (лаговая переменная, которая показывает количество клиентов в день за 7 дней до соответствующей даты), Whole_orders_per_day_lag7 (лаговая переменная, которая показывает количество выданных заказов в день за 7 дней до соответствующей даты).

Для прогнозирования было использовано несколько методов: Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Neural Network Regression и Poisson Regression. На выходе получены метрики качества по 6 моделям (табл. 4).

Таблица 4

Метрики качества полученных предиктивных моделей объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод линейной регрессии, метод байесовской линейной регрессии, метод дерева решений, метод случайного леса, метод нейронных сетей, метод пуассоновской регрессии). Выборка за период 08.09.2021–18.11.2021 поделена на обучающую и тестовую в соотношении 75 на 25 (Split Data: Fraction of rows in the first output dataset = 0,75, not Randomized split)

	Linear Regression	Bayesian Linear Regression	Decision Tree Regression	Decision Forest Regression	Neural Network Regression	Poisson Regression
Mean Absolute Error	8722.19186	8939.96046	11297.60205	7768.003526	49068.576923	7129.825072
Root Mean Squared Error	12147.26227	12269.88033	15379.28751	13463.691727	51186.679126	11066.22034
Relative Absolute Error	0.873229	0.895031	1.131068	0.7777	4.912539	0.713808
Relative Squared Error	0.694869	0.708968	1.113829	0.853639	12.338428	0.576693
Coefficient of Determination	0.305131	0.291032	-0.113829	0.146361	-11.338428	0.423307

Составлено автором самостоятельно на основе данных Microsoft Azure Machine Learning Studio

По результатам предиктивного моделирования, наилучшим методом для прогнозирования объема продаж в пунктах выдачи заказов Wildberries является Poisson Regression, которая показывает хорошие метрики качества относительно прочих моделей. Тем не менее, коэффициент детерминации по-прежнему низок, так как R2 = 0,423.

Эксперимент 4

Поскольку в эксперименте №3 тестовая выборка содержит неагрегированные данные по датам и каждой дате соответствует только одно предсказанное значение переменной Bought_per_day, то каждая метрика качества занижена за счет дубликата анализа проскоренных (от англ. score; здесь и далее — предсказанных) значений. Ввиду этого следующим этапом была очищена тестовая выборка за период 04.11.21–14.11.21 гг., объем которой сократился с 780 до 11 наблюдений. В данном эксперименте в качестве признаков учтем только лаговые переменные. Таким образом, прогнозирование объема продаж на наборе данных с 6 атрибутами

(Date, Day_of_the_week, Bought_per_day, Bought_per_day_lag7, Clients_per_day_lag7, Whole_orders_per_day_lag7) показало следующие результаты (табл. 5).

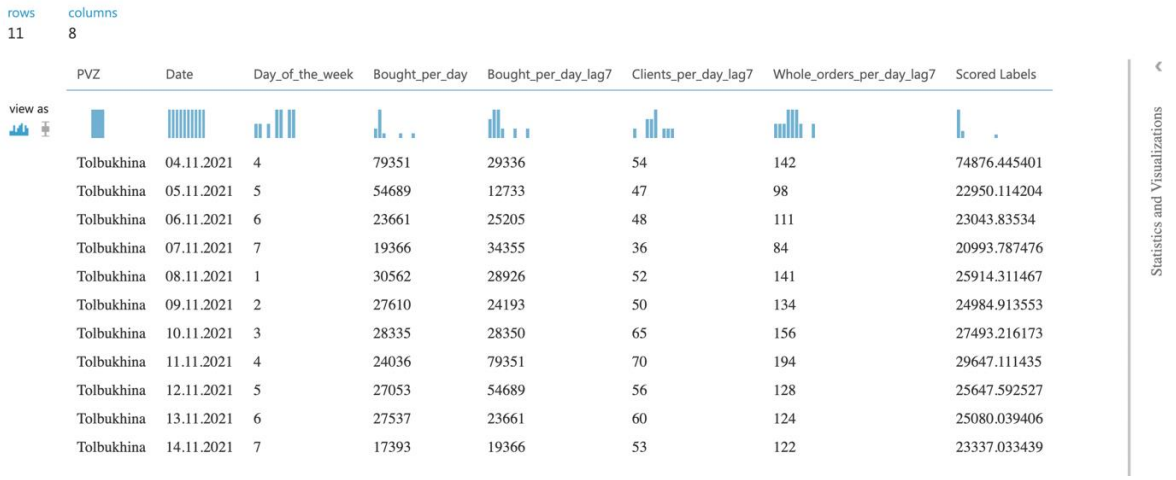
Таблица 5

Метрики качества полученных предиктивных моделей объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод линейной регрессии, метод байесовской линейной регрессии, метод дерева решений, метод случайного леса, метод нейронных сетей, метод пуассоновской регрессии). Выборка за период 08.09.2021–18.11.2021 гг. поделена на обучающую и тестовую в соотношении 75 на 25 (Split Data: Fraction of rows in the first output dataset = 0,75, not Randomized split).

	Linear Regression	Bayesian Linear Regression	Decision Tree Regression	Decision Forest Regression	Neural Network Regression	Poisson Regression
Mean Absolute Error	6742.812051	7156.777742	10469.766246	6735.909091	46660.727273	5635.496753
Root Mean Squared Error	10096.18057	10270.81755	15553.917014	13082.80863	49805.28889	10155.328708
Relative Absolute Error	0.540136	0.573297	0.838686	0.539583	3.737781	0.451434
Relative Squared Error	0.336031	0.347757	0.797526	0.564245	8.177423	0.33998
Coefficient of Determination	0.663969	0.652243	0.202474	0.435755	-7.177423	0.66002

Составлено автором самостоятельно на основе данных Microsoft Azure Machine Learning Studio

Фрагмент проскоренной модели представлен на рисунке 5.


rows 11 columns 8

PVZ	Date	Day_of_the_week	Bought_per_day	Bought_per_day_lag7	Clients_per_day_lag7	Whole_orders_per_day_lag7	Scored Labels
Tolbukhina	04.11.2021	4	79351	29336	54	142	74876.445401
Tolbukhina	05.11.2021	5	54689	12733	47	98	22950.114204
Tolbukhina	06.11.2021	6	23661	25205	48	111	23043.83534
Tolbukhina	07.11.2021	7	19366	34355	36	84	20993.787476
Tolbukhina	08.11.2021	1	30562	28926	52	141	25914.311467
Tolbukhina	09.11.2021	2	27610	24193	50	134	24984.913553
Tolbukhina	10.11.2021	3	28335	28350	65	156	27493.216173
Tolbukhina	11.11.2021	4	24036	79351	70	194	29647.111435
Tolbukhina	12.11.2021	5	27053	54689	56	128	25647.592527
Tolbukhina	13.11.2021	6	27537	23661	60	124	25080.039406
Tolbukhina	14.11.2021	7	17393	19366	53	122	23337.033439

Рисунок 5. *Фрагмент предиктивной модели объема продаж в пунктах выдачи заказов Wildberries на основе применения метода линейной регрессии (составлено автором самостоятельно)*

Лучшие результаты прогнозирования объема продаж в пунктах выдачи заказов Wildberries показал метод линейной регрессии, коэффициент детерминации которой составил 0,664. Тем не менее, для построения более точной модели возможно требуется введение новых атрибутов и учет иных, ранее не принятых к рассмотрению признаков, что является предметом следующих исследований.

Эксперимент 5. Оценка применимости модели спустя год после нововведений Wildberries

К началу 2022 года Wildberries отменила программы субсидирования, которые были актуальны осенью 2021, снизились проценты с оборота, получаемые владельцем ПВЗ на старте.

52ECVN622
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Многие предприниматели жалуются на ужесточение штрафных санкций со стороны Wildberries. Кроме того, с каждым месяцем становится все сложнее найти свободную локацию с хорошим охватом, а отыскав такую, собственник ПВЗ может столкнуться с тем, что через некий промежуток времени по соседству откроется такой же конкурентный пункт, поскольку охранной зоны у ПВЗ Wildberries нет.

Для покупателей в начале 2022 года Wildberries ввел платный возврат товаров. Такое нововведение было призвано повысить ответственность покупателей при оформлении доставки и уменьшить случаи злоупотреблений бесплатными сервисами. Платный возврат для категорий покупателей с низким процентом выкупа, скорее всего, отражается на их потребительском поведении в сторону уменьшения количества заказываемых товаров и увеличения процента выкупа заказа, что, в свою очередь, влияет на выручку собственников пунктов выдачи заказов Wildberries.

Таким образом, в условиях постоянно меняющихся правил торговли, требований к открытию пунктов выдачи заказов и снижения процентов от оборота ПВЗ, который является выручкой собственника точки, автор исследования предполагает, что предложенная предиктивная модель, построенная на данных периода с 08.09.2021 по 18.11.2021 гг., может стать неработоспособной.

Для проверки данной гипотезы необходимо протестировать полученную модель в эксперименте 4 на свежих данных. Для этого автором в качестве тестовой выборки был использован еще один набор неагрегированных по датам данных за период 08.09.2022–14.09.2022 гг. Прогнозная модель построена в студии машинного обучения Microsoft Azure Machine Learning Studio (рис. 6).

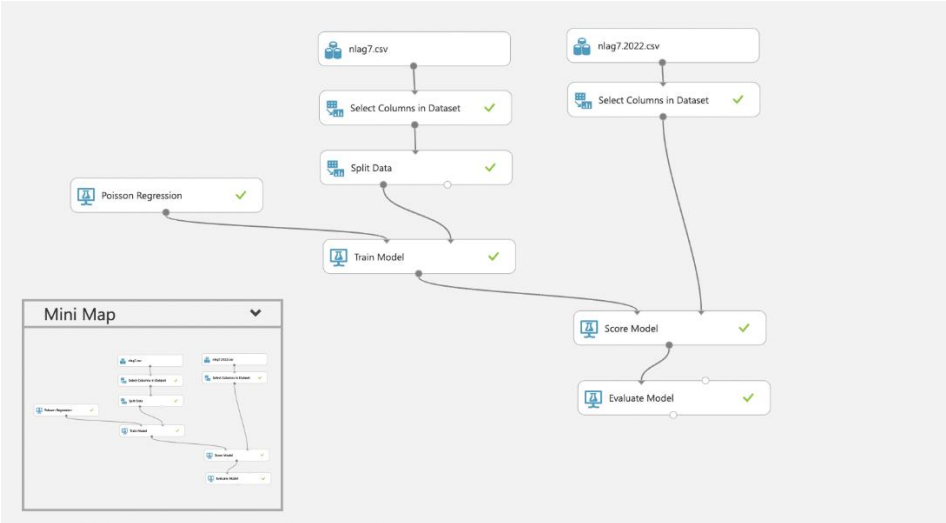


Рисунок 6. *Тестирование полученной модели машинного обучения на новых данных за период 08.09.2022–14.09.2022 гг. (метод — пуассоновская регрессия) для оценки работоспособности модели в условиях изменившихся правил возврата товаров на маркетплейсе Wildberries. Nlag7.cvv — набор данных объемом 3119 наблюдений с 6 атрибутами: Date, Day_of_the_week, Bought_per_day, Bought_per_day_lag7, Clients_per_day_lag7, Whole_orders_per_day_lag7 за период 08.09.2021-18.11.2021 гг. (Split Data: Fraction of rows in the first output dataset = 0,75), nlag7.2022.cvv — набор данных за период 08.09.2022–14.09.2022 гг. (составлено автором самостоятельно)*

Результаты тестирования модели отражены в таблице 6.

52ECVN622
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Таблица 6

Метрики качества полученной в эксперименте 4 предиктивной моделей объема продаж в пунктах выдачи заказов Wildberries на основе алгоритма машинного обучения (метод пуассоновской регрессии), протестированной на новых данных за период 08.09.2022–14.09.2022 гг.

	Метод Poisson Regression
Mean Absolute Error	10249.547385
Root Mean Squared Error	11771.400952
Relative Absolute Error	1.266404
Relative Squared Error	1.688406
Coefficient of Determination	-0.688406

Составлено автором самостоятельно на основе данных Microsoft Azure Machine Learning Studio

Таким образом, мы видим, что перечисленные внешние и внутренние факторы сильно повлияли на качество предиктивной модели объема продаж в пунктах выдачи заказов Wildberries, в результате чего автор делает выводы о необходимости ее доработки на фоне изменяющейся конъюнктуры рынка, что в рамках анализируемой проблематики служит поводом для проведения дальнейших исследований.

Выводы

В рамках исследовательской работы автором приведена описательная статистика по набору данных по пункту выдачи заказов Wildberries, расположенному по адресу г. Москва, ул. Толбухина 13к1, за период 01.09.2021–18.11.2021 объемом 3329 наблюдений, экспериментально построено и протестировано 5 предиктивных моделей объема продаж с различной вариацией используемых признаков и применением 6 методов машинного обучения (Linear Regression, Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Neural Network Regression, Poisson Regression), которые были оценены с помощью 5 метрик качества (Coefficient of Determination, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Relative Squared Error).

Результатом исследования выступает прогнозная модель объема продаж, которая на фоне влияния как внутренних, так и внешних факторов, подробно описанных в настоящей работе, характеризуется снижением адекватности и работоспособности. Апробация наилучшей из построенных 5 предиктивных модели на новых данных доказало необходимость ее доработки в новых реалиях, что является поводом для дальнейших исследований.

ЛИТЕРАТУРА

1. Осин А.А., Фомин А.К., Сологуб Г.Б., Виноградов В.И. Использование методов машинного обучения для решения задач прогнозирования спроса на новый товар в интернет-маркетплейсе // Моделирование и анализ данных. — 2020. — Т. 10. — № 4. — С. 41–50.
2. Артемьев Е.А., Мокшин В.В. Прогнозирование продаж в ритейле на основе методов машинного обучения // Информатика: проблемы, методы, технологии. — 2022. — С. 887–895.

52ECVN622
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

3. Мамиев О.А., Финогенов Н.А., Сологуб Г.Б. Использование методов машинного обучения для решения задач прогнозирования суммы и вероятности покупки на основе данных электронной коммерции // Моделирование и анализ данных. — 2020. — Т. 10. — №. 4. — С. 31–40.
4. Каипов И.К., Чигвинцев К.А. Обзор методов машинного обучения для краткосрочного прогнозирования продаж в обувном ритейле // Веб-программирование и интернет-технологии WebConf2021. — 2021. С. 102–104.
5. Сердинская Ю.А., Мокшин В.В. Использование методов машинного обучения для оценки прогнозирования продаж товара // Информатика: проблемы, методы, технологии. — 2022. — С. 1062–1068.
6. Инюцина В.С., Новиков В.Э. Использование искусственного интеллекта для прогнозирования продаж в сетевой розничной торговле // Логистика и управление цепями поставок. — 2021. — № 2–3. — С. 37–43.
7. Сергеева И.И., Шестова, К.Ю. Цифровая трансформация бизнес-процессов взаимодействия с клиентами // Инфраструктура цифрового развития образования и бизнеса. — 2021. — С. 77–82.
8. Алейникова Ю.Д. Искусственный интеллект в ритейле как фактор повышения конкурентоспособности // Синтез науки и образования в решении глобальных проблем современности. — 2020. — С. 130–132.
9. Охримук Е.С., Размочаева Н.В. Исследование контролируемых алгоритмов решения задачи прогнозирования динамики розничной торговли // Наука настоящего и будущего. — 2020. — Т. 1. — С. 236–239.
10. Sahoko Kaji & Teruo Nakatsuma & Masahiro Fukuhara (ed.), 2021. "The Economics of Fintech," Springer Books, Springer, number 978-981-33-4913-1, December.

Anastasia Atrokhova Nikolaevna
Financial University under the Government of the Russian Federation, Moscow, Russia
E-mail: sserenityy@mail.ru
RSCI: https://elibrary.ru/author_profile.asp?id=1032653

Machine learning methods for predicting sales at the Wildberries order pick-up point

Abstract. A large part of the modern economy is based on information, therefore, various types of e-commerce are becoming very popular and are gradually replacing physical commerce, and online trading platforms with a developed network of points of issue of orders are becoming more common.

Russian marketplaces offer to open a pick-up point as a partner. An entrepreneur who wants to use this option and create such a business needs to use predictive analytics, since most of the large marketplaces, such as Wildberries, transfer partners on a percentage of turnover, moving away from a flat rate for issuance. Thus, the revenue of the owner of the point of issue of orders depends on the volume of sales at a particular point, which makes it important to predict the future volume of sales.

The object of the study is a method for forecasting sales in the retail sector using machine learning algorithms, and the subject is automated tools for forecasting sales volume at Wildberries pick-up points.

The author presents model quality metrics for tested 6 machine learning methods: Linear Regression, Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Neural Network Regression, Poisson Regression.

The result of this study is a predictive model of the sales volume of the Wildberries pick-up point, independently developed by the author. The practical significance of the work is due to the possibility of using this model to predict the revenue of the entrepreneur-owner of the point of issue of orders in order to further compare the predictive data with the item of expenditure and analyze the profitability of the current point of issue of orders.

The main difference and novelty of this work is that previously none of the studies was considered from the standpoint of the interests of the owner of the point of issue of orders.

Keywords: sales forecasting; predictive analytics; machine learning methods; regression; data processing; data analysis; order pick-up point; retail