ЭКОНОМИКА И МАТЕМАТИЧЕСКИЕ МЕТОДЫ, 2025, том 61, № 1, с. 25–37

ТЕОРЕТИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ

Распространенные ошибки использования машинного обучения при прогнозировании событий и новый подход на основе моделей механизмов образования событий

Ю.А. Кораблев,
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Москва; e-mail: yura-korablyov@yandex.ru

В.А. Судаков,
Федеральный исследовательский центр «Институт прикладной математики им. М.В. Келдыша» РАН, Москва; e-mail: sudakov@ws-dss.com

Поступила в редакцию 09.04.2024

Аннотация. Обсуждаются распространенные ошибки, допускаемые исследователями при прогнозировании событий с помощью моделей на основе машинного обучения. Такими ошибками являются: потеря самих событий, вследствие конструирования абстрактных признаков; обучение моделей происходит по клиентам, а не по событиям от клиентов; конструирование искусственных признаков; неправильная валидация и ошибочные метрики качества модели; используются статичные параметры. Приведен разбор совершенных ошибок одного примера с Kaggle. Площадь под ROC-кривой у такого примера очень высокая — 0,88. Однако эта метрика качества рассчитана некорректно. После исправления всех ошибок корректная метрика оказалась 0,599. Представлен иной подход к анализу и прогнозированию событий, который значительно отличается от классических методов машинного обучения. Метод основан на рассмотрении индивидуальных механизмов образования событий для каждого клиента. Строятся модели таких механизмов. Математическими методами восстанавливаются параметры моделей этих механизмов образования событий. Параметры экстраполируются на будущее. Прогноз будущего события получается в результате функционирования модели механизма с установленными значениями параметров. Метрика качества модели, площадь под кривой ROC, составила 0,615, что немного больше, чем в рассматриваемом примере с Kaggle, основанном на машинном обучении. Тем самым показано, что предложенный подход является конкурентным для передовых методов машинного обучения.

Ключевые слова: анализ событий, прогноз событий, машинное обучение, ошибки моделей, механизм образования событий, восстановление параметров, сплайновая коллокация, сглаживающий сплайн, монотонный сплайн, качество прогноза, валидация.

Классификация JEL: C1, C15, C4, C5, C53.

УДК: 330.4; 51-77; 004.9.

Для цитирования: **Кораблев Ю.А., Судаков В.А.** (2025). Распространенные ошибки использования машинного обучения при прогнозировании событий и новый подход на основе моделей механизмов образования событий // *Экономика и математические методы*. Т. 61. № 1. С. 25–37. DOI: 10.31857/S0424738825010039

1. ВВЕДЕНИЕ

Способность прогнозировать события, особенно экономические, невозможно переоценить. Поэтому методы, позволяющие анализировать и прогнозировать экономические события, представляют особый интерес. В данной статье будут рассмотрены основные подходы прогнозирования событий с помощью методов машинного обучения. Будут рассмотрены распространенные ошибки, допускаемые во время применения методов машинного обучения. Оказывается, что такие классические подходы, как методы регрессии и классификации, следует применять с особой осторожностью. При этом всегда остается большая доля неточности, так как такие методы обладают существенными недостатками при прогнозе дискретных событий. Будут рассмотрены соответствующие недостатки. На одном существующем наглядном практическом примере (Nagesh,

2022), опубликованном на Kaggle[1], который собрал чуть ли ни все ошибки, будет показано, как целесообразно подходить к анализу и прогнозированию событий с помощью машинного обучения.

Примечательно, что никто не заметил ошибок в первоначальном варианте решения (Nagesh, 2022). Данное решение получило положительные отзывы на Kaggle, хотя применять машинное обучение таким образом при прогнозе событий в корне неверно (разработчик модели входит в тысячу лучших из 56 тысяч зарегистрированных разработчиков). Поэтому цель данной статьи — обратить внимание научного сообщества и практиков машинного обучения на проблемы применения методов этого машинного обучения для прогнозирования событий является вполне актуальной. Более того, в конце исследования будет предложен оригинальный подход анализа и прогнозирования событий, отличающийся от известных методов прогнозирования на основе машинного обучения. Будет показано, что такой подход превосходит методы машинного обучения (на рассматриваемом примере после исправления всех ошибок). Новый подход к анализу событий извлекает и использует больше информации из набора данных, чем привычные методы[2].

2. ПОСТАНОВКА ЗАДАЧИ

Под событием понимается некоторый факт экономической, социальной или политической жизни, являющийся результатом проявления процессов, происходящих в той или иной сфере. В теории случайных процессов события обычно возникают через случайные интервалы времени. Однако случайность — лишь отражение незнания или непонимания происходящих процессов или неспособность просчитать их будущее появление (споры о существовании истинной случайности, космическом детерминизме или принципе неопределенности оставим философам). Можно также вспомнить определение из теории вероятностей, где событие определяется как множество из возможных подмножеств элементарных исходов, что, по сути, можно интерпретировать как наступление определенной комбинации условий (само выполнение условий и есть событие). По нашему мнению, для экономических событий условия складываются не в результате случайного опыта или испытания. Можно отслеживать каждый необходимый показатель, пытаться анализировать его, выдвигать предположения, гипотезы, строить модели его изменения и тем самым понять и предсказать его будущие значения, после чего проверять условия наступления события. В предлагаемом в конце статьи новом подходе именно так и делается. Далее мы

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15 CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME7'' SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
...	...	...	...	...	...	...	...	...
944463	575312	22083	PAPER CHAIN KIT RETROSPECT	6	2011-11-09 12:49:00	2.95	13588.0	United Kingdom
944464	575312	23355	HOT WATER BOTTLE KEEP CALM	4	2011-11-09 12:49:00	4.95	13588.0	United Kingdom
944465	575312	22110	BIRD HOUSE HOT WATER BOTTLE	6	2011-11-09 12:49:00	2.55	13588.0	United Kingdom
944466	575312	22037	ROBOT BIRTHDAY CARD	12	2011-11-09 12:49:00	0.42	13588.0	United Kingdom
944467	575312	21790	VINTAGE SNAP CARDS	12	2011-11-09 12:49:00	0.85	13588.0	United Kingdom

Рис. 1. Набор данных критикуемого примера

[1] Популярная платформа для организации конкурсов и общения специалистов по обработке данных и машинному обучению.

[2] Программный код, данные, примеры и результаты расчетов опубликованы в открытом репозитории на GitHub по адресу https://github.com/YuAKorablev/Events_prediction

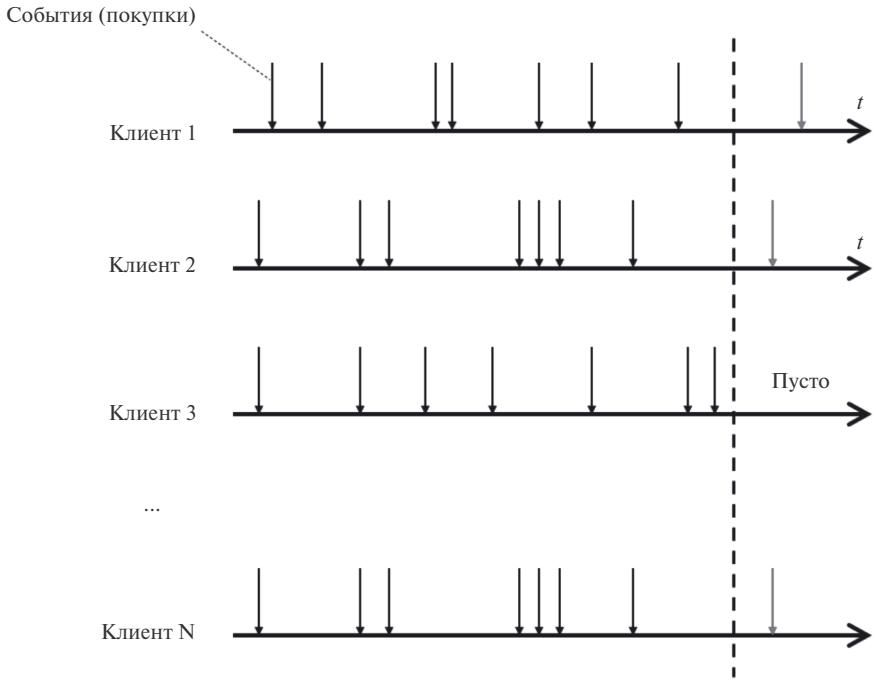

Рис. 2. Задача прогнозирования событий

рассмотрим методы машинного обучения при прогнозировании событий и обсудим распространенные ошибки и проблемы.

Задача из критикуемого примера (Nagesh, 2022) заключается в определении того, кто из клиентов некоторого онлайн-магазина будет совершать покупку в следующем периоде. То есть необходимо предсказать факт будущей покупки каждого клиента. В анализируемом наборе данных — почти миллион транзакций прошлых покупок, содержащих идентификатор клиента, дату, идентификатор товара, количество товара, его цену и другие параметры (рис. 1). По сути, перед автором этого примера стоит задача предсказания будущей покупки (рис. 2). Автор решает ее методами классификации. С методов классификации начнем обсуждать проблемы машинного обучения при прогнозировании событий.

3. ПРОБЛЕМЫ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Методы машинного обучения анализируют события с точки зрения наблюдаемых сопутствующих признаков, выявляются статистические закономерности между наблюдаемыми значениями признаков и некоторым индикатором, сигнализирующим о возникновении события. Именно с ним и связана первая проблема.

3.1. События преобразуются в абстрактные индикаторы. При применении методов классификации для обозначения самого факта наступления события формируют бинарную переменную, принимающую значение 0 или 1. Но для этого сначала необходимо определить фиксированный интервал времени, на котором как раз проверяется, возникло событие или нет. Например, если событие произошло в следующем месяце, то ставим единицу, иначе — ноль. Само событие при этом теряется. Событие могло выпасть в первый день этого месяца, а может, — и в последний. Но наблюдение как было единицей, так единицей и останется. При этом сам выбор ширины интервала ничем не регламентируется. Ширина интервала может быть определена через предметную область, когда слишком большой интервал брать бессмысленно, а выбор слишком узкого интервала приводит к тому, что применяемые методы начинают работать крайне плохо, так как будет слишком много нулевых значений. В итоге обычно останавливаются на некотором среднем варианте. Стоит заметить, что когда событие не попало на интервал, то оно на самом деле могло выпасть очень близко к границе интервала (справа или слева от него). Например, если ширина интервала тот же месяц, но событие запоздало лишь на один день, лишь на одну минуту

или наносекунду, все равно индикатором будет оставаться тот же ноль. Но ведь тот же ноль выбирается, когда событие произошло с задержкой и в два, и в шесть месяцев, и в год, и в два, или же оно не произошло вовсе. Информация о моменте возникновения события теряется или же выбрасывается. То, что событие произошло с задержкой в один день или неделю, можно было бы использовать, но оно игнорируется, как будто его не было.

Сюда же можно отнести потерю другой важной информации, которую несет само событие. Ведь помимо момента (даты) события, событие может содержать и, как правило, содержит дополнительную информацию, например то же количество купленных товаров. Но при подготовке набора данных для применения методов машинного обучения ограничиваются лишь датой, преобразованной в индикатор «0» или «1».

3.2. Усреднение по наблюдениям перемешивает информацию. Вторая проблема связана с тем, что практически во всех методах машинного обучения так или иначе заложено усреднение по выборке наблюдений. Например, в методе наименьших квадратов минимизируется сумма квадратов отклонений: $\sum_{\text{По наблюдениям}}\left(Y_i - f\left(X_i\right)\right)^2 \to \min$. Но что является наблюдениями? Очень часто в методах машинного обучения наблюдениями являются клиенты. То есть в обучающих матрице X и столбце Y в строках указываются данные клиентов, совсем разных клиентов, каждый — со своими индивидуальными вкусами и предпочтениями (рис. 3).

Получается, что усреднение происходит по разным клиентам. Наблюдаемые признаки должны объяснять возникновение / невозникновение события, что для одного клиента, что для другого. Разделяющая гиперплоскость (в методах классификации) строится по отклонениям от этой гиперплоскости точек признаков разных клиентов. В реальности клиенты с близкими по значению признаками могут иметь противоположные предпочтения.

Одним из подходов для устранения этой проблемы является выявление и группировка в отдельные наборы данных схожих наблюдений (клиентов). Далее обучение проводится по каждой группе отдельно. Таким образом якобы решается проблема смешивания информации. Но лишь частично. В каждой группе все равно наблюдениями являются хоть и схожие, но разные клиенты. Более того, сокращение размера выборки приводит к снижению репрезентативности и возможному появлению проблемы оценивания моделей с большим числом параметров. Само разделение выборки на группы также является непростой задачей, ошибки на этом этапе будут влиять на дальнейшее исследование групп наблюдений.

3.3. Признаки конструируются искусственно. Зачастую объясняющие признаки (столбцы обучающей матрицы) формируются искусственно. Например, если имелось большое число событий, как на приведенном выше рис. 2, то часто формируется отдельный столбец с частотой этих событий, среднем временем между событиями, с суммарными объемами покупок. Таким образом, множество отдельных произошедших событий сжимаются до нескольких абстрактных характеристик. В такие моменты происходит потеря информации, так как анализируется не вся совокупность событий, а лишь агрегированные абстрактные показатели. Всю последовательность событий не включают в обучающую матрицу, так как иначе строки получились бы разной длины. В итоге сами события просто выбрасываются, а остаются лишь эти абстрактные агрегированные признаки.

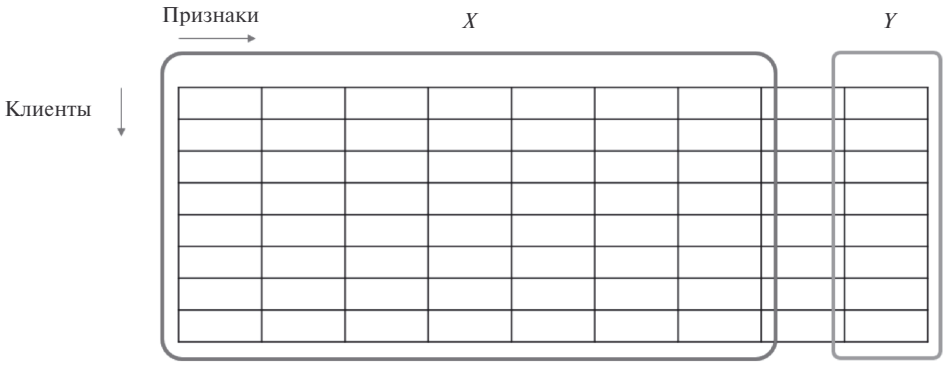

Рис. 3. Матрицы, использующиеся в обучении

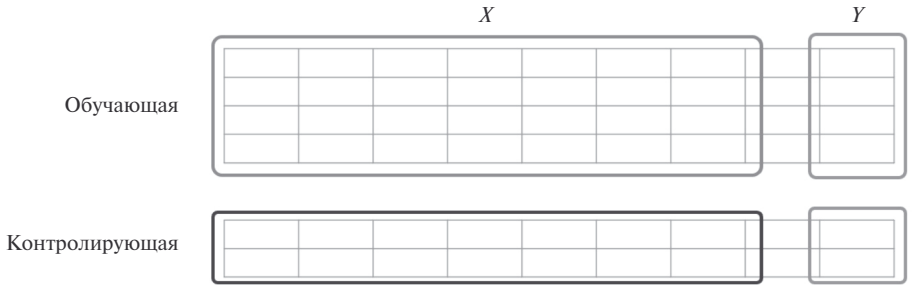

Рис. 4. Стандартный подход при валидации

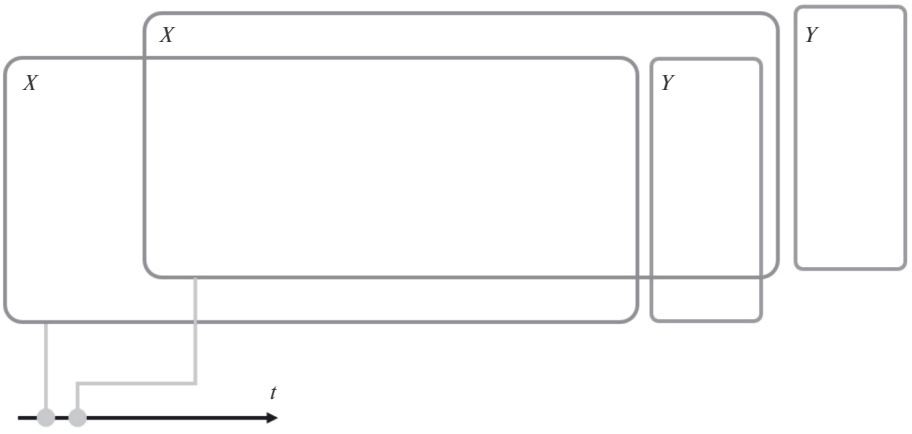

Рис. 5. Правильная валидация

3.4. Неправильная валидация и метрики. Стандартной и рекомендуемой практикой является разбиение имеющейся выборки наблюдений на обучающую и тестирующую[3] (рис. 4). Обычно в обучающую выборку включается большая часть наблюдений, около 80–90%, а в контролирующую — меньшая часть, оставшиеся 10–20%. Но получается, что модель, построенная по данным одних клиентов, должна хорошо объяснять наблюдаемые эндогенные переменные других клиентов. Опускаем описанную выше проблему справедливости такого применения модели. Обратим внимание на то, что валидация в этом случае происходит по наблюдениям, отнесенным к тому же периоду времени, а не к будущим событиям. Наша исходная цель — научиться объяснять и предсказывать будущие события.

Стоит заметить, что в этом случае метрики качества модели могут показывать очень хороший результат, но они будут некорректными. Модель не сможет хорошо предсказывать появления будущих событий.

Валидация обязательно должна проходить по будущим событиям. Для этого следует разделять выборку на обучающую и контролирующую не по наблюдениям (клиентам), и не в рекомендуемых пропорциях (80 на 20%), а по времени и по всем 100% наблюдений (рис. 5). Другими словами, следует подготовить два периода для расчета объясняемой переменной Y, первый — для обучения, а второй — для валидации. При этом обучающую матрицу X следует также подготавливать по выборке наблюдений, в которые не включаются наблюдения из последнего контролирующего периода времени.

[3] Cross-validation: Evaluating estimator performance. SciKit-Learn. Machine Learning in Python (https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation).

3.5. Статичные параметры. Даже если разбиение обучающего и контролирующего набора данных произведено правильно, остается еще одна проблема, связанная с самими методами машинного обучения. Методы машинного обучения получают статические оценки параметров. Следовательно, для модели будут рассчитаны параметры $a_0, a_1, \ldots$ (коэффициенты перед наблюдаемыми экзогенными переменными, признаками). Эти значения коэффициентов будут объяснять ту картину мира, которая запечатлена в обучающей матрице, сформированной в предпоследний момент времени. Но для последнего момента времени в обучающей матрице будет запечатлена уже другая картина мира. Нет никаких гарантий, что в разные моменты времени параметры $a_0, a_1, \ldots$ должны оставаться одними и теми же. С большей уверенностью можно утверждать, что сами параметры должны меняться со временем и быть функцией времени $a_0(t), a_1(t), \ldots$

Может показаться, что включение в обучающую матрицу X признака времени t может решить проблему. На самом деле это не так, параметры $a_0, a_1, \ldots$ останутся статичными. К ним лишь добавится еще один параметр, например a_2 (или несколько параметров), который будет показывать, как наблюдаемый признак времени влияет на объясняемую переменную. Вся временная зависимость будет осуществляться через этот дополнительный параметр. Но одним признаком времени не получится объяснить, как должны были изменяться влияния других признаков на объясняющую переменную.

3.6. Не замечаются скрытые параметры. Если представить задачу прогнозирования события не как задачу классификации, а как регрессионную задачу, когда строится модель не для факта появления события, а для времени возникновения следующего события, одни проблемы исчезают, но некоторые другие проявляются.

Ранее представленная задача (см. рис. 2) может быть интерпретирована таким образом, что по наблюдаемым событиям надо определить зависимость интервалов времени между событиями. В этом случае в качестве объясняющей переменной хорошо взять количество потраченных средств или объем покупки. Можно попробовать объяснить интервалы между покупками объемом совершенной покупки. По-хорошему должна соблюдаться зависимость: чем больше клиент покупает, тем больше интервалы времени между покупками. По крайней мере, этому учат в логистике в моделях управления запасами. Однако это справедливо лишь при стационарном спросе, но не при нестационарном спросе. Для этого легко привести простейший пример из систем управления запасами. Если задать нестационарный спрос, например в виде гармонической функции (рис. 6), задать постоянными максимальный и критический уровни запасов, то наблюдается знакомая пилообразная динамика запасов (рис. 7).

Примечательно, что время между пополнениями запасов сильно разнится более чем в 2 раза, причем в модели не было стохастических факторов. Объем покупок также практически постоянный. Регрессионными методами не определить зависимости разного времени между пополнениями запасов от практически постоянного объема заказа (покупок).

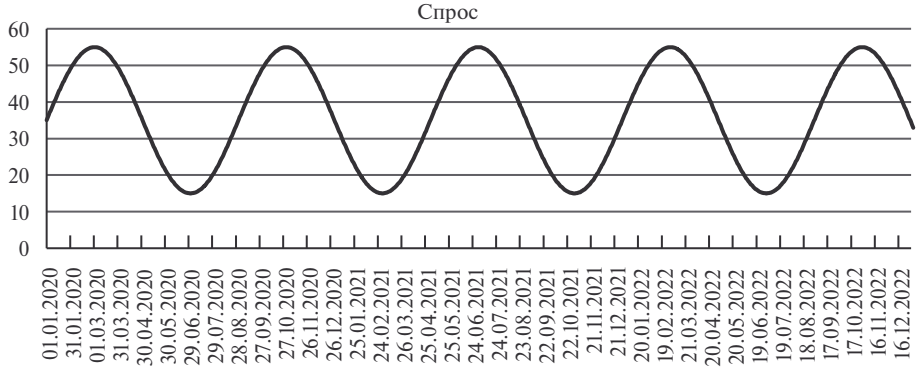

Рис. 6. Нестационарный спрос

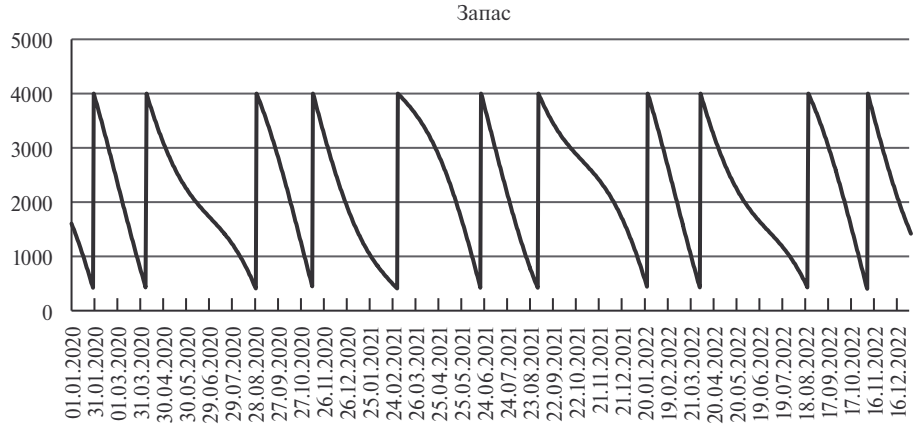

Рис. 7. Динамика запасов

Целесообразно восстанавливать вначале нестационарный спрос, а уже потом получать интервалы между событиями как результат функционирования механизма образования событий.

4. ОШИБКИ КРИТИКУЕМОГО ПРИМЕРА

Рассмотрим типичные ошибки одного примера на Kaggle (Nagesh, 2022). В этом примере на основании набора данных (см. рис. 1), состоящего из транзакций, содержащих идентификатор клиента, дату покупки, идентификатор товара, количество товара, его цену, автор строит классификационную модель, чтобы предсказать, совершит ли клиент онлайн-магазина покупку в следующем периоде. Разработчику модели удалось получить крайне высокую точность (accuracy) 0,92 и площадь под ROC-кривой 0,88. Что может быть лучше, это же — выдающийся результат(!), если бы не ряд ошибок.

1. Самая большая ошибка, даже если не принимать во внимание, что в качестве периода выбрано три следующих месяца — способ формирования эндогенной переменной *Y*. Для этого по каждому клиенту определяется самая последняя покупка до начала следующего квартала и самая первая покупка в следующем квартале. Потом рассчитывается разница в днях между этими двумя покупками (если покупки в следующем квартале не было, то ставится 9999). И если эта разница превышает 90 дней, то объясняемая переменная принимает значение 0 (независимо от того, будет в следующем квартале событие или нет), иначе — 1. Где же здесь подвох? А он состоит в том, что задействуется часть обучающей выборки при формировании объясняемой переменной. В модели как раз используется признак Recency, который показывает, сколько прошло дней от последней покупки обучающего периода до начала следующего квартала. Получается, что есть признак Recency, отчитывающий число дней до начала тестового периода, и одновременно с этим формируется *Y*, отчитывая число дней до первого события тестового периода. Естественно, если Recency будет превышать 90 дней, то *Y* примет значение 0. Таким образом, можно утверждать, что та переменная *Y*, которую объясняет модель, на самом деле конструируется по признаку Recency. Это очень серьезная ошибка, метрики качества модели получаются очень высокими именно из-за этого. Если же исправить эту досадную ошибку[4], которую можно только надеяться, что разработчик модели допустил случайно, и формировать *Y* правильно, проверяя, попало ли событие в следующий квартал

[4] Другим способом исправления ошибки является исключение признака Recency, а также признака RecencyCluster и корректировки признака OverallScore и Segment, которые формируются из суммы кластеров Recency, Frequency, Revenue (исключив влияние Recency на формирование OverallScore и Segment). В результате accuracy (точность) падает до 0,74, а площадь под ROC-кривой — до 0,705. Заметим, что в этом случае получается задача предсказания события в следующем квартале от последнего события каждого клиента. Однако вопрос к корректности результатов все равно остается, так как модель обучается для всех клиентов. Последнее событие у каждого клиента выпадает по-своему. Корректно ли предсказывать прошлый период для одного клиента, обучая модель на будущих периодах других клиентов? Считаем, что все же некорректно.

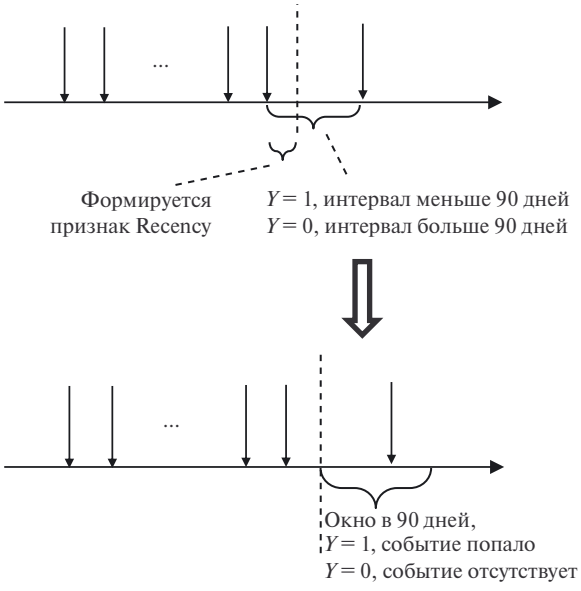

Рис. 8. Ошибка формирования объясняемой переменной Y и ее исправление

или нет, то accuracy падает до 0,76, а площадь под ROC-кривой до 0,73. Визуализация представлена ниже (рис. 8).

2. Другой ошибкой является как раз описанная выше неправильная валидация (см. п. 3.4), когда набор данных разделяется на обучающую и контролирующую выборку (80 и 20% соответственно), относящуюся к одному и тому же периоду времени. Получаются всё еще высокие метрики качества модели. Чтобы исправить эту ошибку, были внесены значительные изменения в открытый программный код на Kaggle. Был реализован правильный способ валидации. Следует правильно организовать разделение выборки на обучающую и контролирующую по времени, без последнего квартала и с последним кварталом, как это показано на рисунке выше (см. рис. 5). Обучающая матрица X сформирована по событиям до предпоследнего квартала, а обучающий столбец Y сформирован по событиям этого предпоследнего квартала. Контролирующая матрица X сформирована по событиям до последнего квартала, а контролирующий столбец Y сформирован по событиям самого последнего квартала. Причем задействуются все 100% наблюдений клиентов (а не 80 и 20%). Наконец, обучив модель по обучающим X и Y, строится прогноз по контролирующей матрице X, который сверяется с контролирующим столбцом Y. В результате самая главная метрика площадью под кривой ROC падает до 0,71 (accuracy остается 0,76). Эти результаты расчета метрики уже корректны. Такой способ валидации единственно правильный, так как перед нами задача предсказания события в будущем, а не задача предсказания события от другого клиента в этом же периоде (в настоящем).

3. Однако одним из главных остается вопрос, насколько справедливо использовать интервал шириной в 90 дней. В обучающем наборе есть клиенты, которые совершают покупки чуть ли не каждые три или семь дней. Насколько обоснованно брать настолько широкий интервал, чтобы прогнозировать будущее событие. В практике динамичной среды реального сектора прогнозы с периодом около месяца более актуальны. Меняя ширину прогнозируемого интервала с 90 дней на 28, получаем падение метрики площади под кривой ROC до 0,614 при неправильной валидации и до 0,605 — при правильной валидации. Выбор ширины интервала в 90 дней формально не является ошибкой, но бесполезен в современном динамичном мире. Возможно даже, что ширина в 28 дней является слишком большой, раз встречаются клиенты, совершающие покупки через 3–7 дней.

4. Еще одна проблема заключается в том, что идентификатор клиента CustomerID включается как признак в обучающую матрицу X. Если его выбросить, то площадь под кривой ROC становится 0,599 (была 0,605) и качество модели немного падает. Эта свидетельствует о том, что идентификатор клиента мог частично объяснять наблюдаемую переменную. Идентификатор CustomerID имеет числовое значение, где-то в модели числовое значение идентификатора одних клиентов объясняет появление событий как тех же, так и других клиентов (обучение всех клиентов). Это кажется нелогичным, вопрос, оставлять или исключать идентификатор клиента из признаков модели, является дискуссионным.

Заметим, что критикуемый пример использует градиентный бустинг (Friedman, 1999, 2001), а именно используется XGBClassifier из библиотеки XGBoost[5]. Такой подход считается одним из самых успешных способов в нахождении скрытых закономерностей между признаками и объясняемой переменной. Недостаток градиентного бустинга состоит в том, что получающиеся модели оказываются неинтерпретируемыми. Достаточно часто для прогнозирования в настоящее время используют рекуррентные нейронные сети (Ехлаков, Судаков, 2022). Они позволяют учесть временную динамику, но проблемы выбора целевых признаков и плохой интерпретируемости остаются такими же, как и для градиентного бустинга.

[5] XGBoost Documentation (https://xgboost.readthedocs.io/en/stable/index.html).

5. ИНОЙ ПОДХОД К ПРОГНОЗИРОВАНИЮ СОБЫТИЙ

Идея иного подхода к анализу и прогнозированию событий заключается в том, что по каждому источнику событий (по клиентам) надо строить индивидуальные модели механизмов образования событий. Применительно к задачам анализа данных, похожим на упомянутую выше задачу, схема подхода приведена ниже (рис. 9).

Идея сильно отличается от стандартного подхода машинного обучения. Не строятся классификационные или регрессионные модели. Вместо этого строится модель механизма образования событий. Основная идея метода описана в монографии (Кораблев, 2023). Подход состоит из пяти этапов: 1) разделение выборки событий по источникам, где они образованы (по клиентам — в текущей задаче этот шаг уже выполнен); 2) выдвижение предположения о механизме образования событий в каждом источнике; 3) восстановление параметров механизма образования событий по выборке событий; 4) экстраполяция параметров этого механизма на будущее любым известным способом; 5) получение прогноза будущих событий, согласно работе механизма образования событий с установленными значениями параметров.

У каждого клиента может быть собственный механизм образования событий. Для начала ограничимся одной моделью. В качестве модели будем использовать модель, похожую на модель управления запасами, только использовать ее будем в обратную сторону. Предполагаем, что клиент совершает покупку тогда, когда его запасы сократились ниже критического уровня. Это и будет механизмом образования событий. Но теперь нам предстоит восстановить параметры этого механизма. Параметром же будет нестационарный спрос. Тут стоит заметить, что по-хорошему такую модель надо строить по каждому товару в отдельности. В нашем случае у нас практически все товары разные. Тогда можно предложить другую интерпретацию: образование событий продиктовано нарастанием потребности что-то купить (шопинга), а потраченные средства удовлетворяют эту потребность. По-другому эту модель можно назвать «деньги жгут карман», чем больше денег потратил клиент, неважно, на какие товары, тем позднее будет следующая покупка. Однако эта модель не регрессионная. Это модель механизма образования событий. Далее мы восстанавливаем параметры этого механизма. Параметром будет являться нестационарная скорость восстановления потраченных средств (скорость нарастания желания совершить покупку) $f(t)$, выраженная в у.е. в день. Вторым параметром будет критический уровень средств, при достижении которого происходит событие — совершение покупки.

Для восстановления параметра скорости восстановления средств $f(t)$ мы приходим к задаче восстановления функции по последовательности определенных интегралов (потраченных средств

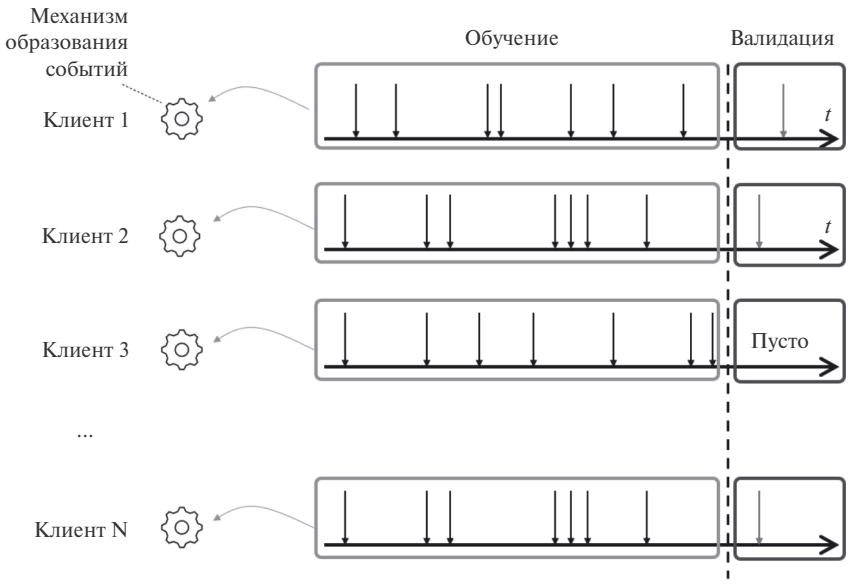

Рис. 9. Схема подхода к анализу и прогнозированию событий

на покупку). Эта задача успешно решается методами сплайновой коллокации, подробности не будем приводить, сами методы и их программная реализация подробно описаны в работах (Korablev, 2022; Кораблев, 2022). Причем восстанавливаемая функция имеет определенный смысл, который предполагает, что функция не должна быть отрицательной. Поэтому восстановление желательно проводить с использованием монотонных сплайнов, когда добавляются условия неотрицательности восстанавливаемой функции. Подробнее о восстановлении неотрицательной функции по разным функционалам с помощью монотонных сплайнов будет описано в следующих работах.

У всех этих методов есть такой параметр, как параметр сглаживания, который через штраф на меру кривизны функции влияет на ее гладкость. Во многих работах по сглаживающим сплайнам, когда функция восстанавливается по зашумленным значениям этой функции, параметр сглаживания принято определять на основе метода L-кривой (Hansen, 1992, 2001) или кросс-валидации (Craven, Wahba, 1978; Golub, Heath, Wahba, 1979). Однако при восстановлении функции по зашумленным функционалам, а конкретно — по наблюдаемым с погрешностью интегралам, метод L-кривой не работает, а метод кросс-валидации работает плохо приблизительно в 50% случаях. Чуть ниже будет описан другой подход по подбору этого гиперпараметра.

Второй параметр механизма, критический уровень средств, при достижении которого происходит покупка, в простейшем случае предполагается стационарным. Для его определения достаточно посчитать средний объем потраченных средств, с корректирующим вычитанием половины значения функции в эту дату $(1/n)\sum_i \left(Y_i - 0{,}5 f(t_i)\right)$, где Y_i — наблюдаемые объемы потраченных средств, $f(t_i)$ — скорость восстановления средств на покупку. Эта корректировка продиктована тем, что обычно в системах управления запасами используются страховые запасы и величина заказа компенсирует их расход. В этом легко убедиться, если моделировать классические модели управления запасами.

После восстановления функции скорости восстановления средств и критического уровня средств начинается четвертый этап — этап экстраполяции параметров. Критический уровень средств предполагаем стационарным. Для экстраполяции нестационарной скорости восстановления средств $f(t)$ можно использовать любой из известных математических подходов. Можно просто переносить динамику прошлого периода, например прошлого года, на следующий период. Можно оценивать разные регрессионные модели, например полиномиальные, чтобы провести экстраполяцию. А можно разложить восстановленную функцию на сумму ограниченного числа гармонических колебаний с помощью алгоритма Куинна–Фернандеса (Quinn, Fernandes, 1991; Quinn, Hannan, 2001). Последний подход кажется более обоснованным, так как низкочастотные гармоники будут показывать долгосрочные тенденции, а высокочастотные гармоники — показывать краткосрочные (частоты и амплитуды определяются алгоритмом).

Однако восстановленный сплайн в самом начале и в самом конце сплайн старается превратиться в линию (вторая производная обращается в ноль в первом и последнем узле по условию натурального сплайна). Из-за этого, а лучше сказать, из-за того, что отсутствуют данные до первого и после последнего наблюдения, на концах сплайн определен хуже всего. Поэтому при разложении восстановленной функции на гармоники желательно пропустить какое-то число наблюдений с самого начала и самого конца. Число откидываемых точек восстановленной функции с левого и правого конца является очередным гиперпараметром, который необходимо как-то задавать.

В итоге получается три гиперпараметра, коэффициент сглаживания и число отбрасываемых слева и справа точек. Для определения этих гиперпараметров разработана отдельная процедура. Идея этой процедуры мотивирована тем, как ищутся гиперпараметры при использовании методов машинного обучения с помощью алгоритмов поиска на сетке. Мы будем искать гиперпараметры с помощью поиска на сетке — для глобальной оптимизации и с помощью алгоритма Нелдера–Мида (Nelder, Mead, 1965) — для локальной оптимизации в окрестности узлов сетки. Но в методах машинного обучения используется метрика качества классификации. У нас будет метрика качества прогноза события.

Прогнозное событие в нашем примере характеризуется двумя показателями, прогнозной датой события t_{n+1} и прогнозным объемом потраченных средств Y_{n+1}. Качество прогноза для будущего события определяем как

$$Score = \left| \frac{\hat{t}_{n+1} - t_{n+1}}{t_{n+1} - t_n} \right| + 0{,}1 \left| \frac{\hat{Y}_{n+1} - Y_{n+1}}{Y_{n+1}} \right|,$$

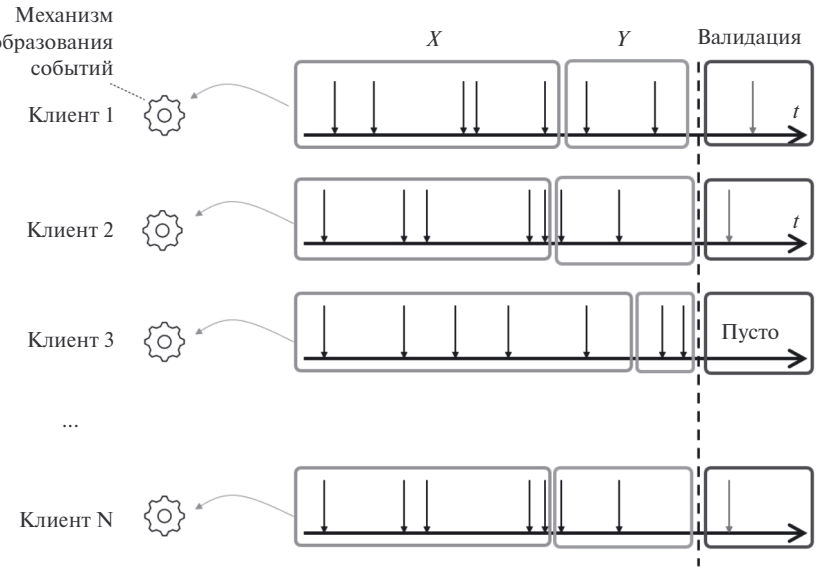

Рис. 10. Схема обучения в новом подходе к анализу событий

где t_{n+1} и Y_{n+1} — реальные значения будущего события, разница $t_{n+1} - t_n$ показывает фактическую ширину интервала между событиями. Тем самым качество прогноза оценивается как относительная погрешность времени до события и соответствующего его объема покупки (с весом 0,1, выбирается эмпирически из соображений, что в первую очередь мы интересуемся датой события).

Конечно же, в качестве последнего события выступает не то событие, по которому происходит валидация. «Окно» с данными должно быть корректно сдвинуто на одну позицию влево.

Однако выяснилось, что подгонка гиперпараметров лишь к одному событию плохо сказывается на способности предсказывать очередное событие. Было решено определять гиперпараметры на основе оценки качества прогноза двух событий. То есть изначально в обучающий набор попадают события без последних трех событий (двух событий обучающего периода и одного события, оставленного для валидации). Для заданных гиперпараметров определяется качество прогноза *Score*1 третьего с конца события. Затем в обучающий набор добавляется реальное третье с конца событие и для тех же заданных гиперпараметров определяется качество прогноза *Score*2 второго с конца события. Их средняя оценка $\left(Score1 + Score2\right)/2$ и выступает в качестве целевой функции во время оптимизации на сетке для подбора гиперпараметров. Последнее же событие используется лишь для валидации построенной модели. Конечно же, для получения прогноза этого последнего события используются выборка без этого последнего события (третье и второе событие с конца включаются). Если проводить аналогию с методами машинного обучения, то обучающий набор X и Y в описанном выше способе нахождения гиперпараметров, а также валидации, изображены на схеме (рис. 10). Пример прогноза события для одного из клиентов показан на (рис. 11).

Рис. 11. Пример прогноза будущего события для одного из клиентов

Примечание. Серая линия показывает среднюю скорость $y_i / (t_{i+1} - t_i)$; сплошная черная линия — восстановленную функцию; пунктирная — экстраполяцию. Треугольник показывает прогноз события, а кружком отмечено реальное событие.

Средняя оценка $(Score1 + Score2)/2$ при оптимальных значениях гиперпараметров может выступать показателем того, насколько хорошо модель механизма образования событий подходит для получения прогноза последнего события. Неправильно ожидать, что механизм образования событий для разных клиентов будет один и тот же, они могут существенно отличаться. Сейчас у нас нет задачи отыскать все такие механизмы для всех клиентов. Наша задача — показать, что такой подход работает и работает не хуже, а иногда даже лучше, чем классическое машинное обучение. Мы предложили одну модель механизма образования событий, посмотрели, для каких клиентов она подходит (у которых оценка погрешности по двум событиям меньше 6%), и для них получили прогноз последнего события. Оказывается, если по прогнозу последнего события опять построить эндогенную переменную со значением 0 или 1 (вернуться в канву методов классификации), то площадь под ROC-кривой становится 0,615 (у метода XGBClassifier для того же набора данных оценка была 0,599, или 0,605, если не выбрасывать признак CustomerID). Иными словами, новый подход дает немного лучший результат, чем самый передовой метод машинного обучения, — правда, для определенных клиентов. В открытом репозитории на GitHub[6] размещен программный код и примеры данных и результатов расчетов.

На самом деле, как было сказано выше, такая метрика все равно приблизительная. Получается та же проблема: хоть событие и не произошло на указанном временном интервале, но оно могло произойти с задержкой всего на один день. Эта метрика не учитывает точности прогноза. Для прогноза последнего события тоже можно было рассчитать качество прогноза *Score* (но только если истинное последнее событие возникло).

6. ЗАКЛЮЧЕНИЕ

Таким образом, мы показали, что при анализе и прогнозировании событий следует крайне осторожно применять методы машинного обучения. Можно совершить множество ошибок, при этом метрики качества модели ничего не покажут, наоборот — они покажут отличный результат. Был проанализирован пример с Kaggle, где были совершены многие из разобранных ошибок.

Был показан пример нового подхода, в котором происходит восстановление механизма образования событий. В данном примере модель механизма можно назвать «деньги жгут карман», так как предполагается, что клиент совершает покупку в онлайн-магазине, когда его накапливаемые средства на покупку превышают заданный критический уровень. Для каждого клиента отдельно оцениваются параметры модели механизма образования событий (также оцениваются гиперпараметры метода). Оценивается качество прогноза на основе третьего и второго с конца события. Для клиентов, у которых качество прогноза хорошее (погрешность прогноза меньше 6%), рассчитывается прогноз последнего события, по которому происходит валидация. Площадь под кривой ROC оказывается 0,615, а в модели на основе XGBClassifier после исправления всех ошибок площадь под кривой ROC была 0,599. Подход на основе составления моделей механизмов образования событий оказывается чуть лучше, чем передовой метод машинного обучения, основанного на градиентном бустинге. Тем самым показано, что такой подход может являться конкурентом передовому методу машинного обучения.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Ехлаков Р.С., Судаков В.А. (2022). Прогнозирование стоимости котировок при помощи LSTM и GRU сетей // *Препринты ИПМ им. М.В. Келдыша*. № 17. 13 с. DOI: 10.20948/prepr-2022-17 [**Ekhlakov R.S., Sudakov V.A.** (2022). Forecasting the cost of quotes using LSTM & GRU networks. *Preprints of IAM after M.V. Keldysh*, 17. 13 p. (in Russian).]

Кораблев Ю.А. (2022). Об одном алгоритме восстановления функции по разным функционалам для прогнозирования редких событий в экономике // *Финансы: теория и практика*. № 3 (26). С. 196–225. DOI: 10.26794/2587-5671-2022-26-3-196-225 [**Korablev Yu.A.** (2022). An algorithm for restoring a function from different functionals for predicting rare events in the economy. *Finance: Theory and Practice*, 3 (26), 196–225 (in Russian).]

Кораблев Ю.А. (2023). *Емкостный метод анализа и прогнозирования редких событий в экономике: монография*. М.: РУСАЙНС. 296 с. ISBN: 978-5-466-04159 [**Korablev Yu.A.** (2023). *Capacity method of analysis and forecasting of rare events in the economy*. Moscow: RUSCIENS. 256 p. (in Russian).]

Craven P., Wahba G. (1978). Smoothing noisy data with spline functions — estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31 (4), 377–403. DOI: 10.1007/BF01404567

[6] https://github.com/YuAKorablev/Events_prediction

Friedman J. (1999). *Greedy function approximation: A gradient boosting machine*. Technical Report. Deptartment of Statistics. Stanford University.

Friedman J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 5 (29), 1189–1232. DOI: 10.1214/aos/1013203451

Golub G.H., Heath M., Wahba G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* 21 (2), 215–223. DOI: 10.1080/00401706.1979.10489751

Hansen P.C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34 (4), 561–580. DOI: 10.1137/1034115

Hansen P.C. (2001). The L-curve and its use in the numerical treatment of inverse problems. In: P. Johnston (ed.). *Computational inverse problems in electrocardiology*. Advances in Computational Bioengineering. Southampton: WIT Press.

Korablev Yu.A. (2022). Restoration of function by integrals with cubic integral smoothing spline in R. *ACM Transactions on Mathematical Software*, 48 (2), 1–17. DOI: 10.1145/3519384 ISSN: 0098-3500

Nagesh S.C. (2022). *Predict customers probable purchase. Kaggle*. Available at: https://www.kaggle.com/code/nageshsingh/predict-customers-probable-purchase

Nelder J.A., Mead R. (1965). A simplex method for function minimization. *The Computer Journal*, 4 (7), 308–313. DOI: 10.1093/comjnl/7.4.308

Quinn B.G., Fernandes J.M. (1991). A fast efficient technique for the estimation of frequency. *Biometrika*, 3 (78), 489–497.

Quinn B.G., Hannan E.J. (2001). *The estimation and tracking of frequency*. Cambridge: Cambridge University Press. 278 p.

Common mistakes in using machine learning when forecasting events and a new approach based on models of the event formation mechanisms

Yu.A. Korablev,
Financial University under the Government of the Russian Federation, Moscow, Russia; e-mail: yura-korablyov@yandex.ru

V.A. Sudakov,
Keldysh Institute of Applied Mathematics of Russian Academy of Sciences (KIAM RAS), Moscow, Russia; e-mail: sudakov@ws-dss.com

Received 09.04.2024

Abstract. The main mistakes made by researchers when predicting events using models based on machine learning are discussed. Such errors are: loss of events themselves, due to the construction of abstract features; models are trained on customers rather than events from customers; construction of artificial features; incorrect validation and erroneous model quality metrics; and static parameters are used. An analysis of the mistakes made in one example from Kaggle is provided. The area under the ROC curve for this example is very high — 0.88, but this quality metric is calculated incorrectly. After correcting all errors, the correct metric turned out to be 0.599. A different approach to analyzing and predicting events is presented, which differs significantly from classical machine learning methods. The method is based on consideration of individual mechanisms of event formation for each client. Mechanism models are being built. Using mathematical methods, the parameters of the models of these event formation mechanisms are restored. Parameters are extrapolated to the future. The forecast of a future event is obtained as a result of the functioning of the mechanism model with established parameter values. The model quality metric, the area under the ROC curve, turned out to be 0.615, which is slightly higher than in the Kaggle example, based on machine learning. Thereby, it is shown that the proposed approach is competitive to advanced machine learning techniques.

Keywords: event analysis, event forecast, machine learning, model errors, mechanism of event formation, parameter recovery, spline collocation, smoothing spline, monotonic spline, forecast quality, validation.

JEL Classification: C1, C15, C4, C5, C53.

UDC: 330.4; 51–77; 004.9.

For reference: **Korablev Yu.A., Sudakov V.A.** (2025). Common mistakes in using machine learning when forecasting events and a new approach based on models of the event formation mechanisms. *Economics and Mathematical Methods*, 61, 1, 25–37. DOI: 10.31857/S0424738825010039 (in Russian).