DOI: 10.17323/2587-814X.2023.1.7.17

Прогнозирование оттока клиентов на основе паттернов изменения их поведения

Ю.А. Зеленков [a]
E-mail: yzelenkov@hse.ru

А.С. Сучкова [b]
E-mail: assuchkova_1@edu.hse.ru

[a] Национальный исследовательский университет «Высшая школа экономики»
Адрес: Россия, 101000, г. Москва, ул. Мясницкая, д. 20

[b] Национальный исследовательский университет «Высшая школа экономики», НИУ ВШЭ – Санкт Петербург
Адрес: Россия, 190121, г. Санкт-Петербург, ул. Союза Печатников, д. 16

Аннотация

Удержание клиентов является одной из главных задача бизнеса, при этом крайне важно распределить ресурсы на удержание в соответствии с потенциальной прибыльностью потребителя. Чаще всего задача прогнозирования оттока клиентов решается на основе RFM (Recency, Frequency, Monetary) модели. В работе предлагается способ расширения RFM модели с помощью оценок вероятности изменения поведения клиента. На основе анализа данных о 33918 покупателях крупной российской торговой сети за 2019–2020 гг. показано, что существуют повторяющиеся паттерны изменения их поведения длительностью в один год. Информация об этих паттернах используется для вычисления необходимых оценок вероятности. Включение этих данных в предиктивную модель на основе логистической регрессии увеличивает точность прогнозирования более чем на 10% по метрикам AUC и геометрическое среднее. Показано также, что данный подход имеет ограничения, связанные с нарушением паттернов поведения в случае внешних шоков, таких как локдаун из-за пандемии COVID-19 в апреле 2020 г. В работе также предложен способ идентификации этих шоков, позволяющий спрогнозировать снижение предиктивной способности модели.

Ключевые слова: отток клиентов, предсказание оттока клиентов, RFM модель, расширение RFM модели, паттерны поведения клиента, предиктивная аналитика

Цитирование: Зеленков Ю.А., Сучкова А.С. Прогнозирование оттока клиентов на основе паттернов изменения их поведения // Бизнес-информатика. 2023. Т. 17. № 1. С. 7–14. DOI: 10.17323/2587-814X.2023.1.7.17

Введение

Концепция управления взаимоотношениями с клиентами (Customer Relationship Management, CRM) подразумевает приобретение и удержание наиболее прибыльных покупателей на основе понимания их ценностей и мотивов, определяющих поведение [1]. При этом издержки на удержание значительно ниже, чем на привлечение новых потребителей, а потеря клиента означает потерю всех покупок, которые он мог бы совершить в течение жизненного цикла [2]. Поскольку не все клиенты одинаково привлекательны для компании в финансовом отношении, крайне важно сначала определить их прибыльность, а затем адекватно распределить ресурсы на их удержание [3].

Задача оптимизации расходов на удержание клиентов решается в два этапа: первый – это сегментация покупателей, а второй – предсказание изменения их поведения. Для сегментации обычно используются методы кластеризации, позволяющие разбить множество клиентов на внутренне однородные группы (классы), которые в то же время сильно различаются между собой [4]. Целью моделей, решающих вторую проблему, является идентификация клиентов, которые могут изменить принадлежность к группе, например, перейти из класса активных покупателей в класс с низкими затратами на покупки [5]. Такой подход получил название прогнозирование оттока клиентов (customer churn prediction). Эту задачу можно свести к проблеме бинарной классификации: используя данные о клиентах за периоды $1, ..., t$, обучить классификатор h, предсказывающий вероятность того, что в период $t + 1$ покупатель останется в той же группе, перейдет в группу, генерирующую больший доход (метка класса 0), или перейдет в группу с совокупно меньшим доходом (метка класса 1). На основании данных прогнозов компании разрабатывают дифференцированные маркетинговые стратегии, направленные на удержание клиентов, относящихся к классу 1 [2].

Наиболее широко применяемым подходом к анализу поведения клиентов является модель RFM, объединяющая данные о времени, прошедшем с момента последней активности клиента (Recency), количестве его покупок за период t (Frequency) и общем объеме затраченных средств (Monetary) за этот же период [6]. Согласно традиционному подходу, база данных клиентов по каждому из трех измерений RFM разделяется на 5 равных сегментов (квантилей). Топ 20% клиентов получают метку 5, следующие 20% – метку 4 и т.д. В итоге с каждым клиентом ассоциируется метка, содержащая три числа, соответствующие квантилям по измерениям RFM, например, 534 или 231. Таким образом, выделяется 125 групп клиентов, потенциально различающихся поведением. Очевидно, что такой подход имеет недостатки, поскольку он не гарантирует, что выделенные группы, во-первых, внутренне однородны, а во-вторых, сильно отличаются друг от друга. Поэтому в последнее время чаще используются методы кластерного анализа (k-средних, самоорганизующиеся карты Кохонена и другие), позволяющие разделить базу данных клиентов на основе формальных метрик [7].

Многие авторы рассматривают вариации RFM модели, расширяя ее за счет дополнительных измерений, в том числе используя динамику поведения клиентов. Например, в [8] предложена модель, учитывающую дисконты, а в [9] учитывается длительность нахождения клиента в определенном кластере.

В контексте задачи прогнозирования оттока клиентов, обнаружение паттернов, описывающих устойчивые траектории перемещения потребителей между кластерами, является одним из важнейших направлений исследований. Для этого используются различные динамические модели, основанные на идентификации паттернов с помощью кластеризации [10, 11], ассоциативных правилах [12] и скрытых марковских моделях [13].

Целью настоящего исследования является разработка метода, позволяющего предсказывать изменение поведения клиентов (т.е. их перемещение из одного класса в другой) с учетом информации о потоках потребителей между группами, накопленной за предыдущие периоды. Согласно нашей гипотезе, интенсивность перехода клиентов из одного класса в другой варьируется в течение года, но существует устойчивый шаблон, повторяющийся из года в год. Таким образом, мы можем рассматривать наблюдаемые частоты переходов клиентов из класса в класс как оценки вероятности того, что клиент покинет кластер, в котором он находится в настоящий момент. Включение этой информации в предиктивную модель должно значительно повысить точность предсказания.

1. Данные и формализация задачи

Для проверки выдвинутой гипотезы был использован набор данных, включающий информацию о покупках, совершенных клиентами крупной российской торговой сети в 2019–2020 годах. Данные были консолидированы помесячно, для каждого покупателя вычислены следующие значения:

- R – количество дней, прошедших с момента последней покупки до начала текущего месяца,
- F – количество покупок в текущем месяце,
- M – общая сумма затрат клиента (доход торговой сети) в текущем месяце.

Другие данные, часто используемые в задачах прогнозирования оттока клиентов (возраст, пол, семейное положение и т.д.), в данном случае не использовались, поскольку они содержат значительное количество пропусков и недостоверных значений. После удаления выбросов и неполных данных был получен набор, содержащий сведения о 33918 клиентах, совершивших не менее пяти покупок за два исследуемых года.

На следующем этапе были выделены устойчивые кластеры пользователей. Для этого данные были разделены на периоды длительностью в один месяц и сгруппированы следующим образом. Клиенты, не совершавшие покупок в рассматриваемый месяц (очевидно, что для них $F = 0$ и $M = 0$), считались принадлежащими к одному кластеру и исключались из анализируемого набора данных, клиенты со значениями $F > 0$ и $M > 0$ сегментировались с помощью алгоритма k-средних. Качество полученного разбиения для различного числа кластеров оценивалось с помощью метрики силуэт [14]. Максимальные значения были получены для разбиения на 3 кластера (один, включающий клиентов без покупок в текущем месяце, и два, выделенных алгоритмом k-средних среди клиентов с покупками). Среднее значение силуэта для 24 месяцев 0,708, минимальное 0,649 для декабря 2019 года и максимальное 0,779 для апреля 2020.

На *рис. 1(a)* представлено изменение среднего значения дохода торговой сети от клиента $\hat{E}[M]$ и соответствующий 95% доверительный интервал для двух кластеров покупателей, совершающих

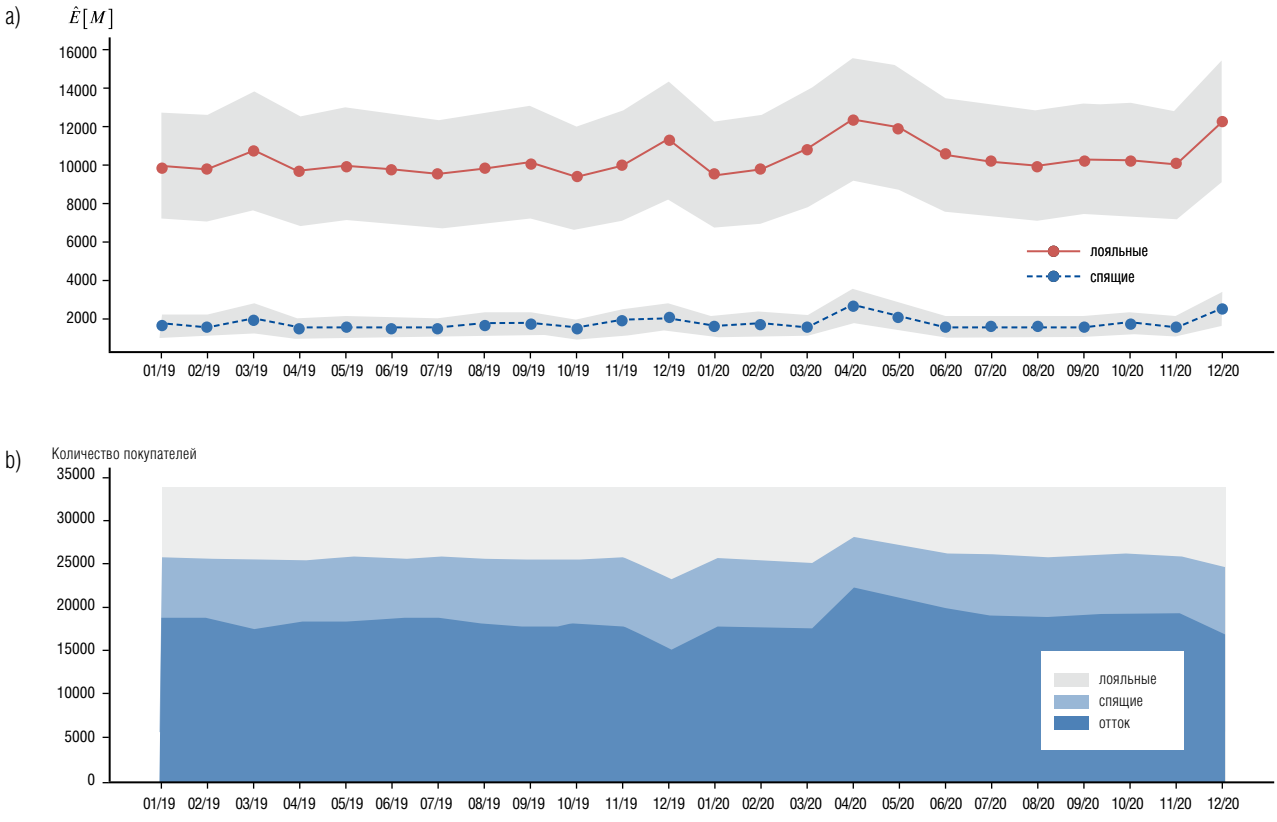

Рис. 1. (a) – средний доход и 95% доверительный интервал для кластеров клиентов, совершающих покупки; (b) – изменение численности клиентов по кластерам.

покупки. Кластер, дающий более высокий средний доход, объединяет покупателей, которых можно определить как «лояльные» (loyal), кластер клиентов, генерирующих меньший доход, можно условно назвать «спящие» (sleeping). Третий кластер, объединяющий клиентов без покупок, мы назвали «отток» (churn). На *рис. 1(b)* представлено изменение численности кластеров.

Анализируя представленные графики, можно сделать вывод о наличии определенных закономерностей, повторяющихся из года в год. Например, в декабре одновременно увеличивается размер кластера loyal и возрастает генерируемый им доход, что связано с сезонными праздниками (Новый год и соответствующие каникулы).

Также был выполнен анализ значимости экзогенных переменных, используя мтод, основанный на измерении снижения точности модели при перемешивании значений интересующего атрибута (permutation importance [15]). В качестве базового классификатора использовалась логистическая регрессия, поскольку эта модель устойчива к возмущениям, а точность модели оценивалась при помощи площади под кривой рабочей характеристики приемника (AUC), так как эта метрика нечувствительна к дисбалансу классов [16]. Полученные результаты показывают, что наиболее значимыми признаками (в порядке уменьшения значимости) являются R_t, M_t, F_t, предшествующие прогнозируемому периоду $t + 1$, т.е. исследуемый процесс изменения поведения клиентов является Марковским. Данные более ранних периодов не влияют на качество предсказания. Это согласуется с результатами других исследователей [4, 17, 18].

Таким образом, формулировка задачи предсказания оттока клиентов может быть уточнена следующим образом. Пусть X – множество описаний клиентов, а $Y = \{0, 1\}$ – множество меток классов. Требуется построить алгоритм h: $X \to Y$, способный классифицировать произвольный объект $x \in X$ по известной конечной обучающей выборке $D = \{[R_t, M_t, F_t]_1, y_1, ..., [R_t, M_t, F_t]_m, y_m\}$, где $y \in Y$, а вектор $(R_t, M_t, F_t) \in X$ – признаковое описание объекта.

2. Моделирование потоков клиентов

Как уже отмечалось выше, многие исследователи фокусируются на расширении вектора признаков X, включая в него дополнительные признаки, что повышает точность работы алгоритма классификации. В данной работе предлагается использовать информацию о динамике потоков клиентов между кластерами.

Данная идея инспирирована эпидемиологическими моделями (ЭМ), которые рассматривают перемещение людей между различными группами: инфицированные, заболевшие, выздоровевшие и т.д. [19]. Интенсивность перемещения из группы A в группу B определяется коэффициентом α^{AB}, который фактически определяет долю членов группы A, перешедших в B. В большинстве ЭМ эти коэффициенты рассматриваются как экспоненциально распределенные случайные величины, однако в нашем случае можно вычислить их точные значения, поскольку численность всех групп известна во все моменты времени. Второе отличие заключается в том, что ЭМ обычно предполагают ограниченное число возможных траекторий перемещения между группами, в нашем же случае клиент может перемещаться из своей группы в любую другую. Учитывая все сказанное, динамику клиентов можно представить следующими разностными уравнениями:

$$L_{t+1} = L_t + \alpha^{SL}_{t+1} S_t + \alpha^{CL}_{t+1} C_t - [\alpha^{LS}_{t+1} + \alpha^{LC}_{t+1}] L_t$$

$$S_{t+1} = S_t + \alpha^{LS}_{t+1} L_t + \alpha^{CS}_{t+1} C_t - [\alpha^{SL}_{t+1} + \alpha^{SC}_{t+1}] S_t$$

$$C_{t+1} = C_t + \alpha^{SC}_{t+1} S_t + \alpha^{LC}_{t+1} L_t - [\alpha^{CL}_{t+1} + \alpha^{CS}_{t+1}] C_t.$$

Здесь L_t, S_t, C_t – численность клиентов в кластерах «лояльные» (loyal), «спящие» (sleeping) и «отток» (churn) соответственно в момент времени t;

α^{AB}_{t+1} – коэффициент потока, определяющий, какая часть клиентов, находящихся в момент времени t в группе A перейдет в период $t + 1$в группу B.

Индекс $t + 1$ в данном случае означает, что значение этого коэффициента станет известно только после наступления момента $t + 1$.

Коэффициент α^{AB}_{t+1} может быть вычислен как $\alpha^{AB}_{t+1} = F^{AB}_{t+1} / A_t$, где F^{AB}_{t+1} – количество клиентов (поток), перешедших из группы A в B в интервал между t и $t + 1$; A_t – количество клиентов в группе A в момент t.

Сравнительный анализ коэффициентов потоков представлен на *рис. 2–4*. Верхние графики на каждом рисунке представляют значения коэффициентов по месяцам 2019 и 2020 года (в данном случае нижний индекс переменной означает год), а также разность между ними. Области, заштрихованные различными оттенками, представляют различ-

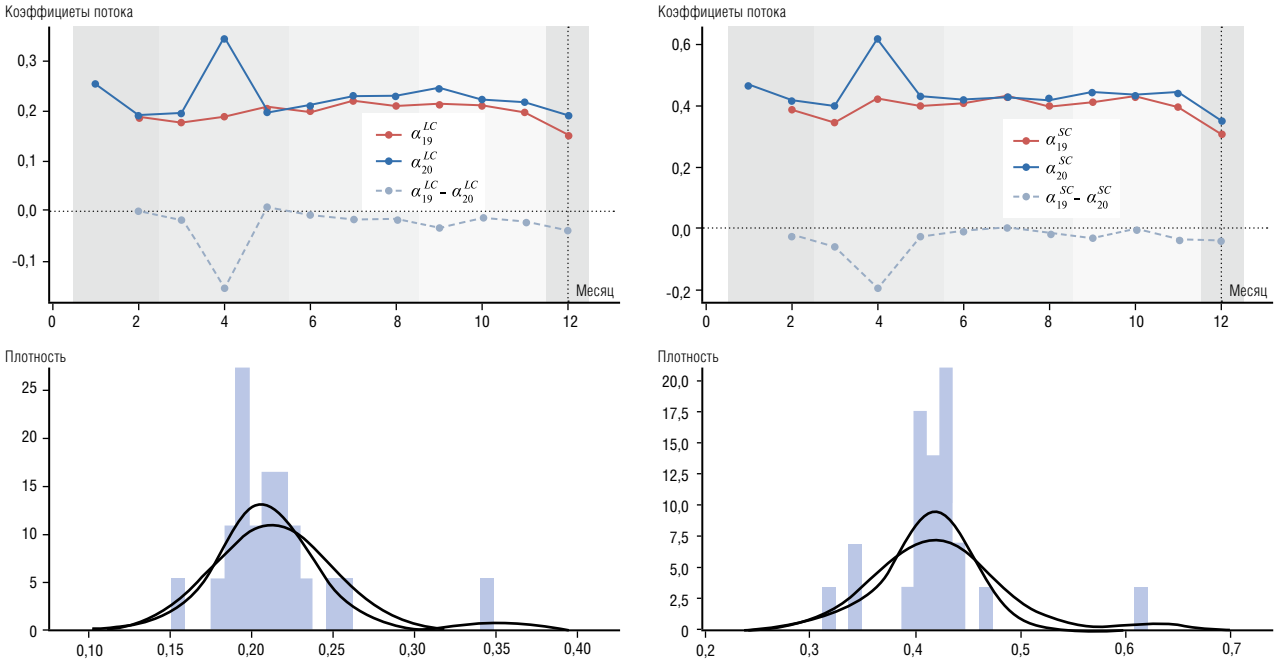

Рис. 2. Коэффициенты потоков в кластер C (churn).

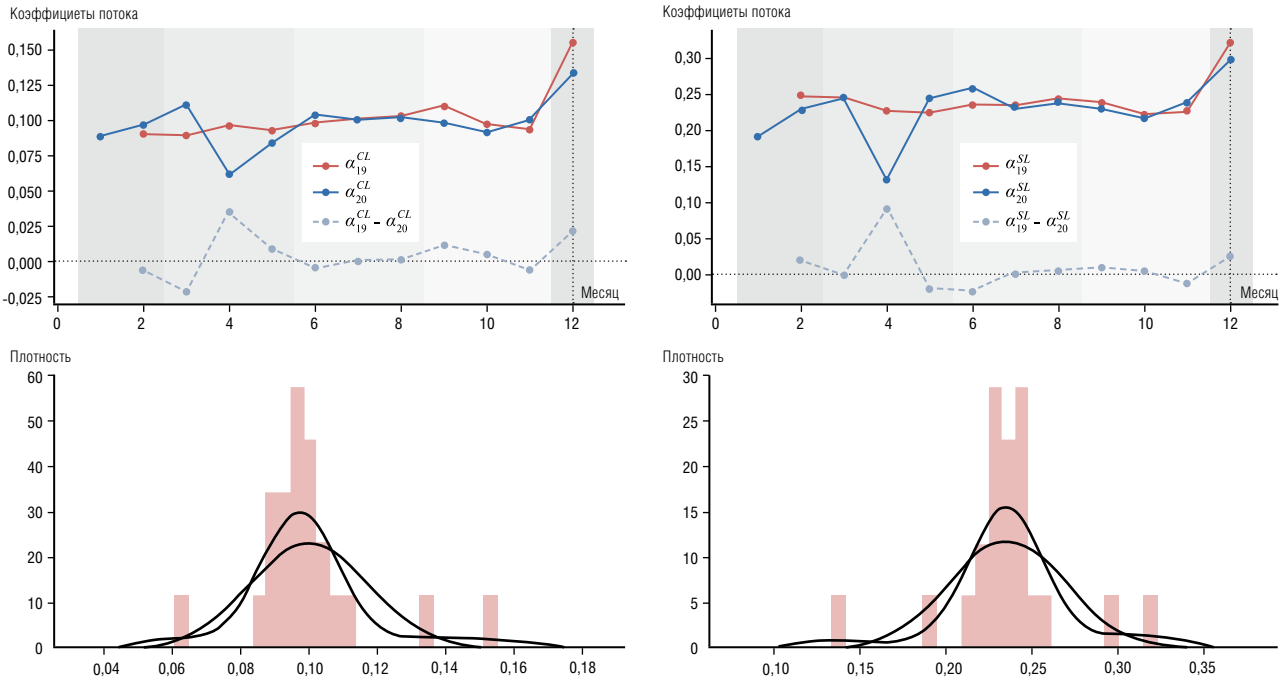

Рис. 3. Коэффициенты потоков в кластер L (loyal).

ные сезоны (зима, весна, лето и осень). Видно, что сезонность, связанная со временем года, отсутствует. Вертикальная пунктирная линия соответствует декабрю, когда, как отмечалось выше, наблюдается рост числа покупок. Из этих графиков следует, что значения коэффициентов за разные годы достаточно близки, разница между ними стремится к нулю (отмечен горизонтальной пунктирной линией). Единственное исключение наблюдается в апреле. Это объясняется тем, что в апреле 2020 г. был введен локдаун, связанный с пандемией COVID-19, что привело к снижению активности покупателей.

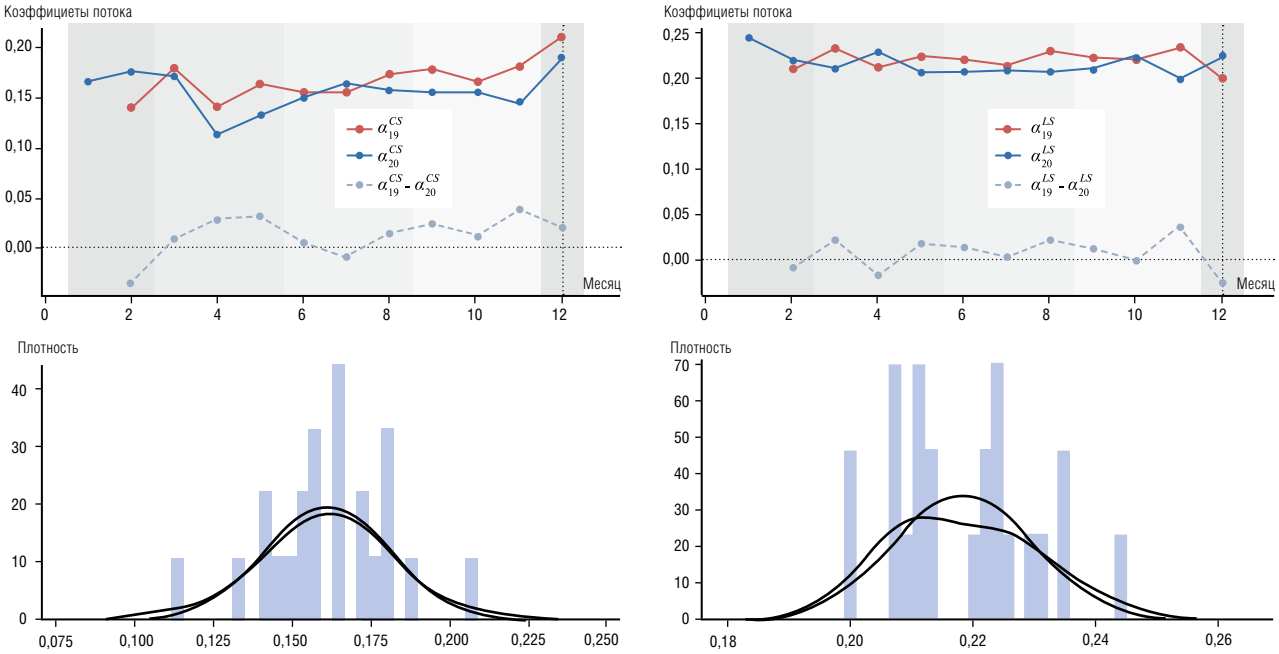

Рис. 4. Коэффициенты потоков в кластер S (sleeping).

Нижняя пара графиков на каждом рисунке представляет распределение значений соответствующих коэффициентов за оба исследуемых года, а также ядерную оценку плотности (kernel density estimation, KDE) этого распределения и нормальное распределение со средним и дисперсией, вычисленным по наблюдаемым значениям. Из этих графиков видно, что распределение значений коэффициентов потоков близко к нормальному. Также на основании представленной информации можно предположить, что временные ряды, представляющие значения коэффициентов за оба года, являются стационарными.

Для проверки предположения о стационарности временных рядов выполнен расширенный тест Дики-Фуллера (Augmented Dickey-Fuller, ADF), значения соответствующей статистики представлены в *таблице 1*, столбец ADF. Полученные результаты показывают, что нулевая гипотеза о наличии единичных корней и, следовательно, нестационарности ряда отвергается для всех коэффициентов кроме α^{CL} и α^{LS}. Однако, если исключить наблюдение, соответствующее апрелю 2020 года, вносящее наибольшие возмущения, то все ряды становятся стационарными (столбец ADF_{-4}).

Кроме того, для оценки подобия коэффициентов за два исследуемых года были вычислены две меры (*таблица 1*): косинусное сходство (cosine similarity, CS) и средняя абсолютная ошибка в процентах (mean absolute percentage error, MAPE)

$$CS(P,N)=\frac{\sum_{i=1}^{n}P_iN_i}{\sqrt{\sum_{i=1}^{n}P_i^2}\sqrt{\sum_{i=1}^{n}N_i^2}},$$

$$MAPE(P,N)=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{P_i-N_i}{P_i}\right|.$$

В данном конкретном случае P и N – вектора коэффициентов потоков за 2019 и 2020 годы соответственно.

Косинусное сходство – это косинус угла между двумя векторами в n-мерном пространстве, где n соответствует количеству значений в последовательности. Как следует из *таблицы 1*, угол между векторами коэффициентов потоков практически равен нулю, т.е. направления векторов совпадают. Относительно высокие значения *MAPE* также объясняются результатами локдауна из-за пандемии COVID-19, это доказывает сравнение столбцов *MAPE* (%) и $MAPE_{-4}$ (%), вычисленных соответственно на всех данных и данных, из которых был исключен апрель.

Таким образом, можно предположить, что включение информации о потоках между группами в число признаков при решении задачи прогнозирования оттока клиентов позволит повысить точность классификации. Используя коэффициенты α, мож-

Таблица 1.

Результаты тестирования на стационарность и меры подобия коэффициентов α за 2019 и 2020 гг.

Коэффициенты		ADF	ADF_{-4}	*CS*	*MAPE* (%)	$MAPE_{-4}$ (%)
Потоки в кластер C	α^{LC}	–5,088*	–4,121**	0,982	16,0	9,2
	α^{SC}	–4,594*	–4,855*	0,993	10,4	6,8
Потоки в кластер L	α^{CL}	–2,777**	–7,968*	0,990	10,9	8,3
	α^{SL}	–5,367**	–6,612*	0,992	8,4	5,0
Потоки в кластер S	α^{CS}	–1,431	–10,757*	0,992	12,4	11,7
	α^{LS}	–0,995	–5,649*	0,997	7,5	7,5

*p <0,01; **p < 0,1.

но вычислить оценку $\hat{p}$ условной вероятности $p(y = 1|X)$ перемещения клиента из группы с высокими затратами на покупки в кластер с меньшими затратами, т.е. из кластера loyal в кластеры sleeping или churn и из кластера sleeping в кластер churn. В соответствии с условиями, сформулированными выше, оценка вероятности того, что клиент относится к классу 1, определяется как

$$\hat{p}_t = \alpha_t^{LS} + \alpha_t^{LC}; \ if\ cluster = loyal$$
$$\hat{p}_t = \alpha_t^{SS}; \quad if\ cluster = sleeping.$$

Поскольку оценка $\hat{p}_{t+1}$ для прогнозируемого периода неизвестна, введем дополнительную переменную $\hat{v}_t$, учитывающий потенциальные изменения $\hat{p}_{t+1}$ на основе данных предыдущего года

$$\hat{v}_t = \hat{p}_{t+1-q} - \hat{p}_{t-q},$$

где q – временной лаг, соответствующий длительности паттерна повторяющегося поведения клиентов. В данном случае $q = 12$.

3. Результаты и дискуссия

Для проверки эффективности предложенного подхода был проведен численный эксперимент по прогнозированию оттока покупателей на основе модели логистической регрессии. Модель обучалась на двух наборах данных, первый включал метрики $D_1 = [R, F, M]$, второй был расширен за счет предложенных здесь переменных $D_2 = [R, F, M, \hat{p}, \hat{v}]$. Задачей модели было по данным в период t определить класс объекта в момент $t + 1$. Обучение, соответственно, производилось на наборе данных $[(R, F, M, \hat{p}, \hat{v})_{t-1}, y_t]$.

Результаты последовательного тестирования модели на валидационных выборках за разные периоды представлены в *таблице 2*. Используемые метрики – площадь под кривой рабочей характеристики приемника (AUC) и геометрическое среднее

$$G_{mean} = \sqrt{TPR \cdot TNR},$$

где *TPR* и *TNR* – доля правильно классифицированных объектов позитивного ($y = 1$) и негативного ($y = 0$) классов соответственно.

Выбор этой метрики дополнительно к AUC обоснован тем, что при прочих равных условиях, геометрическое среднее имеет более высокое значение для сбалансированных предсказаний по обоим классам [16].

Как следует из *таблицы 2*, включение переменных $\hat{p}$, $\hat{v}$ обеспечивает увеличение точности предиктивной модели (в большинстве случаев более 10%) в периоды, когда отсутствуют внешние шоки. Таким образом, предложенный подход может быть использован для разработки индивидуальных стратегий удержания пользователей в относительно стабильное время.

В то же время, модель оказывается бесполезной, когда влияние внешних возмущений является катастрофическим, причем это влияние проявляется с задержкой (см. результаты за период 06/20). В этом случае предиктивная способность модели соответствует случайному угадыванию ($AUC = 0,5$), все прогнозируемые объекты классифицируются как принадлежащие классу 0. Это вполне объяснимо, поскольку для предсказания периода $t = 6$ на основе данных периода $t = 5$ используется мо-

Таблица 2.

Значение метрик *TNR*, *TPR*, *AUC* и G_{mean} для логистической регрессии

Период	$D_1 = [R, F, M]$		$D_2 = [R, F, M, \hat{p}, \hat{v}]$		$\frac{G_{mean}(D_2)}{G_{mean}(D_1)}$	$\frac{AUC(D_2)}{AUC(D_1)}$
	G_{mean}	*AUC*	G_{mean}	*AUC*		
04/20	0,811	0,812	0,856	0,863	1,055	1,063
05/20	0,739	0,745	0,741	0,747	1,004	1,003
06/20	0,781	0,781	0,000	0,500	0,000	0,640
07/20	0,767	0,767	0,846	0,854	1,102	1,113
08/20	0,750	0,750	0,750	0,750	1,000	1,000
09/20	0,733	0,734	0,839	0,850	1,145	1,158
10/20	0,758	0,758	0,841	0,853	1,109	1,125
11/20	0,743	0,743	0,842	0,854	1,134	1,148
12/20	0,742	0,742	0,831	0,839	1,120	1,131

дель, обученная классифицировать объекты в момент $t = 5$ на данных $t = 4$, когда поведение клиентов резко изменилось вследствие локдауна. Это означает, что в апреле 2020 года произошло нарушение паттерна поведения клиентов, на основе которого строятся оценки вероятности изменения поведения. Однако, данная ситуация не является критической, поскольку снижение предиктивной способности модели вследствие внешних шоков является вполне прогнозируемым и может быть учтено при ее использовании.

Сигналом, предупреждающим о потенциальном снижении точности модели, является значительное отклонение в момент t текущих значений коэффициентов α^{AB} от значений, зафиксированных для этого момента в предыдущие годы. Это отклонение может быть обнаружено при анализе графиков, представленных на *рис. 2–4* либо статистическими методами. При наличии такого отклонения для прогноза в $t + 2$ целесообразно использовать модель, обученную на данных $D_1 = [R, F, M]$.

Заключение

В работе показано, что в анализируемой торговой сети наблюдается повторяющийся паттерн перехода клиентов между группами с одинаковым поведением длительностью в один год. Использование информации о потоках клиентов между этими группами позволяет оценить вероятность изменения их поведения. По сравнению с традиционной RFM моделью точность прогнозирования повышается более чем на 10%.

Также продемонстрированы ограничения предлагаемого подхода, которые связаны с нарушением паттерна поведения вследствие внешних шоков. Предложен способ идентификации такого нарушения, что позволяет предсказать деградацию предиктивной способности модели.

В заключение перечислим также возможные направления повышения эффективности предложенного метода:

- Использование выборки, включающей более чем 2 года. Это позволит более точно определить средние значения коэффициентов α^{AB}, а также учесть их тренды.
- Выделение периодов меньшей продолжительности (например, неделя). Потенциально это может позволить обнаружить паттерны изменения поведения меньшей периодичности, что повысит точность прогнозирования в краткосрочной перспективе.
- Использование более сложных моделей, чем логистическая регрессия. ■

Литература

1. Yeh I.C., Yang K.J., Ting T.M. Knowledge discovery on RFM model using Bernoulli sequence // Expert Systems with Applications. 2009. Vol. 36. No. 3. P. 5866–5871. https://doi.org/10.1016/j.eswa.2008.07.018
2. Kotler P., Armstrong G. Principles of Marketing. NY: Pearson Prentice Hall, 2006.
3. Huang S.C., Chang E.C., Wu H.H. A case study of applying data mining techniques in an outfitter's customer value analysis // Expert Systems with Applications. 2009. Vol. 36. No. 3. P. 5909–5915. https://doi.org/10.1016/j.eswa.2008.07.027
4. Wei J.-T., Lee S.-Y., Wu H.-H. A review of the application of RFM model // African Journal of Business Management. 2010. Vol. 4. No. 19. P. 4199–4206. https://doi.org/10.5897/AJBM.9000026
5. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., Chatzisavvas K.C. A comparison of machine learning techniques for customer churn prediction // Simulation Modelling Practice and Theory. 2015. No. 55. P. 1–9. https://doi.org/10.1016/j.simpat.2015.03.003
6. Hughes A.M. Strategic database marketing. NY: Probus Publishing, 1994.
7. Ernawati E., Baharin S.S.K., Kasmin F. A review of data mining methods in RFM-based customer segmentation // Journal of Physics: Conference Series. 2021. Vol. 1869. No. 1. Article 012085. https://doi.org/10.1088/1742-6596/1869/1/012085
8. Heldt R., Silveira C.S., Luce F.B. Predicting customer value per product: From RFM to RFM/P // Journal of Business Research. 2021. No. 127. P. 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001
9. Peker S., Kocyigit A., Eren P.E. LRFMP model for customer segmentation in the grocery retail industry: A case study // Marketing Intelligence and Planning. 2017. Vol. 35. No. 4. P. 544–559. https://doi.org/10.1108/MIP-11-2016-0210
10. Chen Y.L., Kuo M.H., Wu S.Y., Tang K. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data // Electronic Commerce Research and Applications. 2009. Vol. 8. No. 5. 241–251. https://doi.org/10.1016/j.elerap.2009.03.002
11. Hosseini M., Shabani M. New approach to customer segmentation based on changes in customer value // Journal of Marketing Analytics. 2015. Vol. 3. No. 3. P. 110–121. https://doi.org/10.1057/jma.2015.10
12. Akhondzadeh-Noughabi E., Albadvi A. Mining the dominant patterns of customer shifts between segments by using top-k and distinguishing sequential rules // Management Decision. 2015. Vol. 53. No. 9. P. 1976–2003. https://doi.org/10.1108/MD-09-2014-0551
13. Lemmens A., Croux C., Stremersch S. Dynamics in the international market segmentation of new product growth // International Journal of Research in Marketing. 2012. Vol. 29. No. 1. P. 81–92. https://doi.org/10.1016/j.ijresmar.2011.06.003
14. Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. No. 20. P. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
15. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. No. 1. P. 5–32.
16. Zelenkov Y., Volodarskiy N. Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers // Expert Systems with Applications. 2021. No. 185. Article 115559. https://doi.org/10.1016/j.eswa.2021.115559
17. Fader P.S., Hardie B.G., Lee K.L. RFM and CLV: Using iso-value curves for customer base analysis // Journal of Marketing Research. 2005. Vol. 42. No. 4. P. 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415
18. Lumsden S.A., Beldona S., Morrison A.M. Customer value in an all-inclusive travel vacation club: An application of the RFM framework // Journal of Hospitality & Leisure Marketing. 2008. Vol. 16. No. 3. P. 270–285. https://doi.org/10.1080/10507050801946858
19. Bjørnstad O.N., Shea K., Krzywinski M., Altman N. The SEIRS model for infectious disease dynamics // Nature Methods. 2020. No. 17. P. 557–558. https://doi.org/10.1038/s41592-020-0856-2

Об авторах

Зеленков Юрий Александрович

доктор технических наук;

профессор департамента бизнес-информатики, Высшая школа бизнеса, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Сучкова Ангелина Сергеевна

студент бакалавриата, образовательная программа «Международный бизнес и менеджмент», Санкт-Петербургская школа экономики и менеджмента, Национальный исследовательский университет «Высшая школа экономики», НИУ ВШЭ – Санкт-Петербург, 190121, г. Санкт-Петербург, ул. Союза Печатников, д. 16;

E-mail: assuchkova_1@edu.hse.ru

Predicting customer churn based on changes in their behavior patterns

Yury A. Zelenkov [a]
E-mail: yzelenkov@hse.ru

Angelina S. Suchkova [b]
E-mail: assuchkova_1@edu.hse.ru

[a] HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] HSE University – Saint-Petersburg
Address: 16, Soyuza Pechatnikov Street, St. Petersburg 190121, Russia

Abstract

Customer retention is one of the most important tasks of a business, and it is extremely important to allocate retention resources according to the potential profitability of the customer. Most often the problem of predicting customer churn is solved based on the RFM (Recency, Frequency, Monetary) model. This paper proposes a way to extend the RFM model with estimates of the probability of changes in customer behavior. Based on an analysis of data relating to 33 918 clients of a large Russian retailer for 2019–2020, it is shown that there are recurring patterns of change in their behavior over a single year. Information about these patterns is used to calculate the necessary probability estimates. Incorporating these data into a predictive model based on logistic regression increases prediction accuracy by more than 10% on the metrics AUC and geometric mean. It is also shown that this approach has limitations related to the disruption of behavioral patterns by external shocks, such as the lockdown due to the COVID-19 pandemic in April 2020. The paper also proposes a way to identify these shocks, making it possible to forecast degradation in the predictive ability of the model.

Keywords: customer churn, customer churn prediction, RFM model, RFM model extension, customer behavior patterns, predictive analytics

Citation: Zelenkov Y.A., Suchkova A.S. (2023) Predicting customer churn based on changes in their behavior patterns. *Business Informatics*, vol. 17, no. 1, pp. 7–17. DOI: 10.17323/2587-814X.2023.1.7.17

References

1. Yeh I.C., Yang K.J., Ting T.M. (2009) Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, vol. 36(3), pp. 5866–5871. https://doi.org/10.1016/j.eswa.2008.07.018
2. Kotler P., Armstrong G. (2006) *Principles of Marketing*, 11th ed. NY: Pearson Prentice Hall.
3. Huang S.C., Chang E.C., Wu H.H. (2009) A case study of applying data mining techniques in an outfitter's customer value analysis. *Expert Systems with Applications*, vol. 36(3), pp. 5909–5915. https://doi.org/10.1016/j.eswa.2008.07.027
4. Wei J.-T., Lee S.-Y., Wu H.-H. (2010) A review of the application of RFM model. *African Journal of Business Management*, vol. 4(19), pp. 4199–4206. https://doi.org/10.5897/AJBM.9000026
5. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., Chatzisavvas K.C. (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9. https://doi.org/10.1016/j.simpat.2015.03.003
6. Hughes A.M. (1994) *Strategic Database Marketing*. NY: Probus Publishing.

7. Ernawati E., Baharin S.S.K., Kasmin F. (2021) A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, vol. 1869(1), 012085. https://doi.org/10.1088/1742-6596/1869/1/012085

8. Heldt R., Silveira C.S., Luce F.B. (2021) Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, vol. 127, pp. 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

9. Peker S., Kocyigit A., Eren P.E. (2017) LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence and Planning*, vol. 35(4), pp. 544–559. https://doi.org/10.1108/MIP-11-2016-0210

10. Chen Y.L., Kuo M.H., Wu S.Y., Tang K. (2009) Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, vol. 8(5), pp. 241–251. https://doi.org/10.1016/j.elerap.2009.03.002

11. Hosseini M., Shabani M. (2015) New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, vol. 3(3), pp. 110–121. https://doi.org/10.1057/jma.2015.10

12. Akhondzadeh-Noughabi E., Albadvi A. (2015) Mining the dominant patterns of customer shifts between segments by using top-k and distinguishing sequential rules. *Management Decision*, vol. 53(9), pp. 1976–2003. https://doi.org/10.1108/MD-09-2014-0551

13. Lemmens A., Croux C., Stremersch S. (2012) Dynamics in the international market segmentation of new product growth. *International Journal of Research in Marketing*, vol. 29(1), pp. 81–92. https://doi.org/10.1016/j.ijresmar.2011.06.003

14. Rousseeuw P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

15. Breiman L. (2001) Random Forests. *Machine Learning*, vol. 45(1), pp. 5–32.

16. Zelenkov Y., Volodarskiy N. (2021) Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications*, vol. 185, 115559. https://doi.org/10.1016/j.eswa.2021.115559

17. Fader P.S., Hardie B.G., Lee K.L. (2005) RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, vol. 42(4), pp. 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

18. Lumsden S.A., Beldona S., Morrison A.M. (2008) Customer value in an all-inclusive travel vacation club: An application of the RFM framework. *Journal of Hospitality & Leisure Marketing*, vol. 16(3), pp. 270–285. https://doi.org/10.1080/10507050801946858

19. Bjørnstad O.N., Shea K., Krzywinski M., Altman N. (2020) The SEIRS model for infectious disease dynamics. *Nature Methods*, vol. 17, pp. 557–558. https://doi.org/10.1038/s41592-020-0856-2

About the authors

Yury A. Zelenkov

Dr. Sci. (Tech.);

Professor, Department of Business Informatics, Graduate School of Business, HSE University, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Angelina S. Suchkova

Student, BSc Program «International business and management», Saint Petersburg School of Economics and Management, HSE University – Saint Petersburg, 16, Soyuza Pechatnikov Street, St. Petersburg 190121, Russia;

E-mail: assuchkova_1@edu.hse.ru