УДК 004.8:658.8

ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ В РОССИЙСКОМ E-COMMERCE НА ОСНОВЕ NLP-АНАЛИЗА ОТЗЫВОВ И ТРАНЗАКЦИОННЫХ ДАННЫХ

Хвостова А.Н.[1], Варнин А.С. [1]
Научный руководитель – Шалыгина П.М. [1]
[1]Университет ИТМО
nastiahvostova07@mail.ru

Введение

Рынок электронной коммерции в России вступает в фазу жесткой конкуренции за существующего клиента. Стоимость привлечения нового покупателя существенно превышает затраты на удержание текущего, при этом повышение уровня удержания способно значительно увеличить прибыль компании [1]. Однако существующие системы, интегрированные в CRM, до сих пор опираются преимущественно на транзакционные метрики: давность последней покупки, частота транзакций, динамика среднего чека [2]. Они позволяют определить «кто уходит», но не объясняют «почему».

Исследования в области поведенческой экономики показывают, что эмоциональная оценка клиентского опыта является важным фактором лояльности [3]. Однако большинство моделей прогнозирования оттока опираются преимущественно на транзакционные показатели и редко учитывают текстовые данные отзывов [4]. Зарубежные исследования демонстрируют потенциал применения методов NLP для повышения точности прогнозирования [5].

Цель исследования — разработка и эмпирическая проверка модели прогнозирования оттока в российском e-commerce, интегрирующей транзакционные данные и NLP-анализ отзывов.

Основная часть

Эмпирическую базу исследования составляет открытый датасет T-ECD, опубликованный Центром искусственного интеллекта «Т-Технологии» в 2025 г. [6]. Данный набор данных, сформированный на основе анонимизированных цифровых следов пользователей сервисов «Т-Банка», является одним из крупнейших открытых датасетов для исследований электронной коммерции и включает сведения о 135 млрд взаимодействий 44 млн пользователей с 30 млн товаров. Датасет имеет кросс-доменную структуру и объединяет транзакционные данные, поведенческие логи, реакции на промопредложения и тексты отзывов. Период сбора данных варьируется от одного года до 3,5 лет, что обеспечивает репрезентативность поведенческих паттернов [7].

Преимуществом данного источника для задачи прогнозирования оттока является наличие в домене Reviews не только числовых оценок, но и предобученных эмбеддингов текстов отзывов. Для пилотного тестирования и отладки пайплайнов планируется использовать облегчённую версию T-ECD-Small (около 1 млрд взаимодействий) [7].

Методологический дизайн включает четыре этапа. На первом производится операционализация оттока: отсутствие транзакций в течение периода, превышающего критический порог. Второй этап — формирование признакового пространства: транзакционные признаки на основе RFM-метрик [2] и текстовые признаки, полученные путем агрегации эмбеддингов отзывов. Третий этап — моделирование и валидация. В качестве базовых алгоритмов выступают методы градиентного бустинга; для сравнения используются логистическая регрессия и случайный лес. Валидация осуществляется на временных срезах для предотвращения look-ahead bias. Метрики качества — AUC-ROC, precision@k и коэффициент Lift. Четвертый этап — интерпретация результатов с

использованием SHAP для выявления факторов, детерминирующих отток, включая тематические паттерны в отзывах. На основе анализа зарубежных исследований выдвигается гипотеза о том, что включение NLP-признаков обеспечивает статистически значимое повышение качества прогноза на 5–15% по сравнению с моделями, использующими исключительно транзакционные данные [4].

Дизайн исследования. Исследование носит экспериментальный характер. На основе датасета T-ECD формируется выборка пользователей, для которых рассчитываются транзакционные признаки и агрегируются эмбеддинги отзывов. Обучаются модели с различными наборами признаков (транзакционные, текстовые и комбинированные), после чего проводится сравнительный анализ качества прогнозирования и интерпретация результатов с использованием SHAP-анализа.

Выводы

Реализация программы исследования позволит впервые для российского рынка e-commerce количественно оценить вклад текстов отзывов в прогнозирование оттока. Практическая значимость — в интеграции подхода в CRM для раннего выявления клиентов группы риска и персонализации удерживающих коммуникаций, что будет способствовать повышению LTV. Теоретическая значимость — в эмпирическом обосновании необходимости учета эмоционально-оценочных компонентов потребительского опыта в прогностических моделях маркетинга.

Литература

1. Zaghloul M., Barakat S., Rezk A. MNeuralTab: Integrating meta-modeling and neural networks for customer churn prediction in e-commerce // Discover Applied Sciences. 2025. Vol. 7. Article 569. DOI: 10.1007/s42452-025-07157-0.
2. Скорынина И. RFM-анализ вашего поведения в банке // Газпромбанк Tech. 2024. URL: https://www.gazprombank.tech/blog/1190/ (дата обращения: 10.02.2026).
3. Lemon K.N., Verhoef P.C. Understanding Customer Experience Throughout the Customer Journey // Journal of Marketing. 2016. Vol. 80, No. 6. P. 69–96. DOI: 10.1509/jm.15.0420
4. Imani M., Joudaki M., Beikmohammadi A., Arabnia H. Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning // Machine Learning and Knowledge Extraction. 2025. Vol. 7, No. 4. P. 3, P. 33 DOI: 10.3390/make7030105.
5. Amosu O.R., Kumar P., Fadare A.M., Shopade O.F., Faniyi O.O., Adekunle T.A. Data-driven personalized marketing: deep learning in retail and E-commerce // World Journal of Advanced Research and Reviews. 2024. Vol. 23, No. 2. P. 788–796. DOI: 10.30574/wjarr.2024.23.2.2395.
6. T-Tech E-commerce Cross-Domain Dataset (T-ECD) // Hugging Face. 2025. URL: https://huggingface.co/datasets/t-tech/T-ECD (дата обращения: 10.02.2026).
7. Группа «Т-Технологии» опубликовала крупный датасет для рекомендательных систем T-ECD // Компьютерра. 2025. 26 сентября. URL: https://www.computerra.ru/324373/gruppa-t-tehnologii-opublikovala-krupnyj-dataset-dlya-rekomendatelnyh-sistem-t-ecd/ (дата обращения: 12.02.2026).