УДК 004.85:339.13

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Н.А. Бабошин[1], Т.П. Рубцова[1]

[1]Самарский университет, г. Самара
e-mail: baboshinikki90@gmail.com

Аннотация

В условиях роста данных и технологий компании нуждаются в инструментах анализа поведения клиентов. В данной работе представлено исследование подходов машинного обучения для прогнозирования оттока. Особое внимание уделено балансировке данных при работе с несбалансированными выборками. Обучены и интерпретированы разные модели – от классических методов до нейронных сетей.

Ключевые слова: машинное обучение, прогнозирование оттока, потребительское поведение, дисбаланс классов, логистическая регрессия, деревья решений, ансамблевые методы, нейронные сети.

Введение

Предсказание оттока клиентов критически важно для бизнеса: оно помогает выявить закономерности, оценить маркетинговые акции и разработать стратегии удержания аудитории [1]. Методы машинного обучения превосходят традиционные подходы, так как обрабатывают большие объемы данных в реальном времени, обнаруживают сложные взаимосвязи и автоматизируют рутинные задачи.

1. Постановка задачи

Пусть имеется множество клиентов $C = \{c_1, c_2, \dots, c_N\}$, для каждого из которых известен набор признаков $x_i \in \mathbb{R}^d$, характеризующих поведение и взаимодействие клиента с компанией.

Цель состоит в построении функции $f: \mathbb{R}^d \to \{0, 1\}$, которая по признакам клиента x_i предсказывает метку y_i, где

$$y_i = \begin{cases} 1, & \text{если клиент } c_i \text{ склонен к оттоку (уходит)} \\ 0, & \text{если клиент } c_i \text{ остается лояльным} \end{cases}$$

В более общем виде задача сводится к оценке вероятности оттока клиента: $p_i = P(y_i = 1 \mid x_i) = f(x_i)$, где $p_i \in [0, 1]$ – вероятность, что клиент c_i покинет компанию в заданный прогнозный период.

Обучение модели происходит на размеченной выборке $\{(x_i, y_i)\}_{i=1}^{N}$.

2. Подготовка данных

Для построения моделей прогнозирования оттока клиентов были использованы открытые данные, которые охватывают широкий спектр характеристик [2]. При анализе целевой переменной был выявлен сильный дисбаланс классов, что является типичной

ситуацией для задач прогнозирования оттока: доля ушедших клиентов существенно меньше доли оставшихся. Распределение клиентов по оттоку представлено на рисунке 1 в виде круговой диаграммы.

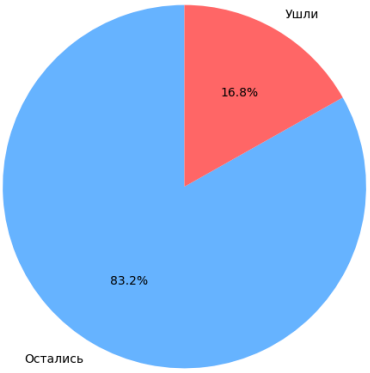


Рисунок 1 – Распределение клиентов по оттоку

Для устранения дисбаланса некоторые модели используют встроенную балансировку через веса классов. Для нейронной сети применялся метод ADASYN, генерирующий синтетические образцы для миноритарного класса с учётом распределения сложных примеров [3].

3. Метрики качества

При прогнозировании оттока с сильным дисбалансом классов accuracy неприменима, поскольку модель может игнорировать отток и всё равно показывать высокий результат. Ключевое внимание следует уделять метрике recall, которая считается по формуле (1). Именно она показывает, насколько полно модель находит ушедших клиентов.

$$Recall = \frac{TP}{TP+FN}, \quad (1)$$

где *TP* (True Positives) – количество ушедших клиентов, верно предсказанных моделью; *FN* (False Negatives) – количество ушедших клиентов, которых модель не распознала.

Дополнительно для оценки качества моделей использовались ROC-кривая и Precision-Recall кривая. Более информативной считается Precision-Recall кривая, которая фокусируется на качестве предсказаний именно для миноритарного класса. Обе кривые представлены на рисунке 2.

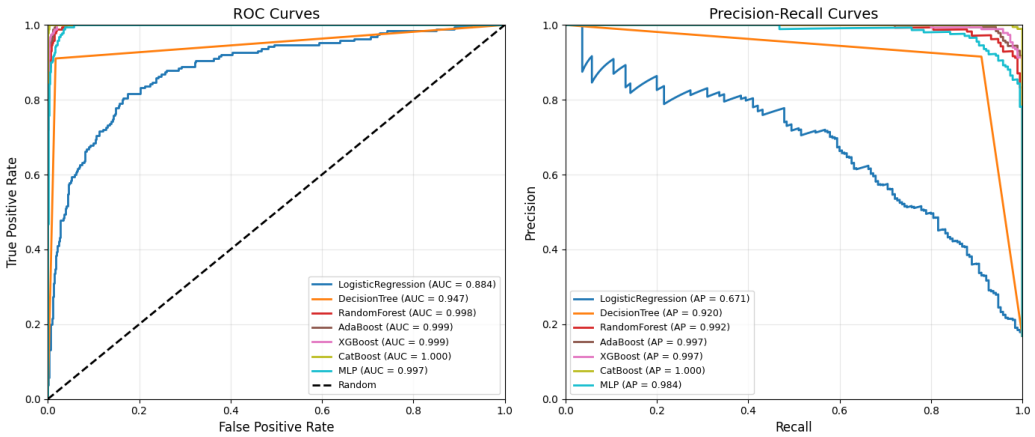


Рисунок 2 – ROC- и Precision-Recall кривые

4. Сравнение моделей

Модели были построены и оценены с использованием Python-библиотек: pandas, numpy, scikit-learn и PyTorch. Гиперпараметры моделей были подобраны с помощью решетчатого поиска (Grid Search). Результаты представлены в таблице 1.

Таблица 1 – Результаты оценки моделей

Модель	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0,7869	0,4324	0,8421	0,5714
Decision Tree	0,9707	0,9153	0,9105	0,9129
Random Forest	0,9840	0,9725	0,9316	0,9516
AdaBoost	0,9876	0,9783	0,9474	0,9626
XGBoost	0,9858	0,9394	0,9789	0,9588
CatBoost	0,9964	0,9794	1,0000	0,9896
Neural Network (MLP)	0,9627	0,8217	0,9947	0,9000

Наилучшие результаты показал CatBoost, обеспечив идеальный баланс точности и полноты. Ансамблевые методы (бустинг, случайный лес) значительно превзошли логистическую регрессию и нейронную сеть, продемонстрировав высокую надёжность и минимальное число ошибок. Градиентный бустинг был признан оптимальным подходом для решения задачи прогнозирования оттока [4].

5. Интерпретация прогнозов модели

Для удержания клиентов компаниям важно понимать причины оттока. Интерпретацию сложных моделей обеспечивает метод SHAP (теория игр), оценивающий вклад признаков [5]. На рисунке 3 (а, б) представлена локальная и глобальная важность признаков по SHAP-значениям.

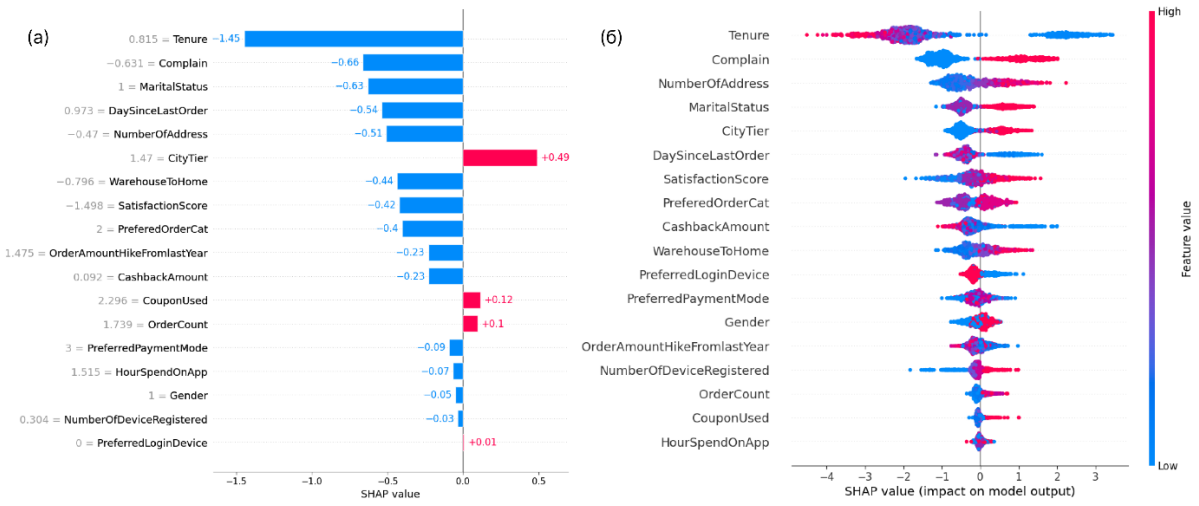


Рисунок 3 – Диаграммы распределения вклада признаков в локальный (а) и глобальный (б) прогноз модели

Заключение

Проведённый анализ показал, что ансамблевые модели превзошли классические методы по точности прогноза оттока. Нейронные сети показали потенциал для дальнейшего улучшения. Создание подобных инструментов полезно для CRM и маркетинговых подразделений: он сокращает расходы на удержание, своевременно

идентифицируя клиентов группы риска [6]. Будущие исследования сосредоточатся на объяснимом искусственном интеллекте, методах для малых выборок и адаптации моделей к поведению клиентов в реальном времени.

Список использованных источников

1. Gold, C. S. Fighting churn with data: The science and strategy of customer retention / Carl S. Gold. – Manning Publications, 2020. – 504 p.

2. Verma, A. Ecommerce customer churn analysis and prediction / A. Verma. – URL: https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction (дата обращения: 29.03.2026).

3. Mohammed, R. Machine learning with oversampling and undersampling techniques: overview study and experimental results / R. Mohammed, J. Rawashdeh, M. Abdullah // 2020 5th International Conference on Information and Communication Systems (ICICS). – IEEE, 2020. – P. 243-248.

4. Bogaert, M. Ensemble methods in customer churn prediction: a comparative analysis of the state-of-the-art / M. Bogaert, L. Delaere // Mathematics. – 2023. – V. 11. – P. 1137.

5. Lundberg, S. M. A Unified approach to interpreting model predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems 30 (NIPS 2017). – Curran Associates, 2017. – P. 4765–4774.

6. E-commerce customer churn prevention using machine learning-based business intelligence strategy / J. Shobana, Ch. Gangadhar, Rakesh Kumar Arora [et al.] // Measurement: Sensors. – 2023. – V. 27, № 100728.

ANALYSIS AND PREDICTION OF CUSTOMER CHURN USING MACHINE LEARNING METHODS

N.A. Baboshin[1], T.P. Rubtsova[1]

[1]Samara University, Samara
e-mail: baboshinikki90@gmail.com

Abstract

With the rapid growth of data and technology, companies require tools to analyze customer behavior. This paper presents a study of machine learning approaches for customer churn prediction. Particular emphasis is placed on data balancing when working with imbalanced datasets. Various models were trained and interpreted, ranging from classical methods to neural networks.

Keywords: machine learning, churn prediction, consumer behavior, class imbalance, logistic regression, decision trees, ensemble methods, neural networks.