

МЕТОДИ ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ В АНАЛІТИЦІ КОРИСТУВАЦЬКИХ ДІЙ У SAAS: ВІД КЛАСИФІКАЦІЇ ДО ПРОГНОЗУВАННЯ

Шаш Максим, Клименко Микита

METHODS OF APPLYING MACHINE LEARNING IN USER BEHAVIOR ANALYTICS
FOR SAAS: FROM CLASSIFICATION TO PREDICTION

Shash Maksym, Klymenko Mykyta

Державний університету інформаційно-комунікаційних технологій, вул. Солом'янська,
7, Київ, 03110, Україна.

Abstract. *The research explores prediction and classification tasks in SaaS platforms by leveraging user behavior analytics. Key challenges such as class imbalance and temporal dynamics are addressed using advanced machine learning techniques, including ensemble models and transformers. The case study presented shows that using categorical data transformation and class imbalance methods can be applied to improve model performance, with a practical case showing an F1-score of 0.84 after feature encoding and class balancing.*

Поведінкові дані користувачів SaaS-платформ та методи їх класифікації і прогнозування є ключовим об'єктом сучасних досліджень у сфері аналітики. Особлива увага приділяється методам обробки, представлення та моделювання поведінкових, демографічних і транзакційних даних з урахуванням дисбалансу класів і часової динаміки, що є предметом цього дослідження. Метою роботи є підвищення точності прогнозування відтоку користувачів за допомогою сучасних методів машинного навчання. SaaS-платформи збирають різноманітні дані — від event logs і session metadata до фінансової інформації та звернень у службу підтримки — що дозволяє глибоко аналізувати взаємодії користувачів, покращувати продукти, підвищувати лояльність і сприяти сталому зростанню.

Основні виклики в аналізі поведінки користувачів SaaS — це висока розмірність і розрідженість даних, що ускладнюють моделювання; дисбаланс класів при прогнозуванні рідкісних подій (наприклад, churn); а також часові динаміки, через які поведінка користувачів змінюється, потребуючи адаптивних моделей для підтримки точності прогнозів.

У SaaS-аналітиці класифікація є ключовою для прогнозування churn, сегментації клієнтів і аналізу прийняття функцій. Вона базується на поведінкових, демографічних та транзакційних даних. Логістична регресія популярна через простоту й інтерпретованість, особливо для бінарних задач. Дерева рішень дозволяють моделювати нелінійні взаємодії між ознаками. Ансамблеві методи, такі як Random Forest і XGBoost, підвищують точність за рахунок об'єднання кількох моделей.

Сучасні підходи, як BBE-LSWCM, використовують як довгострокові, так і короткострокові clickstream-дані для кращого прогнозування churn, враховуючи динаміку поведінки користувачів [1].

Дисбаланс класів — поширена проблема, особливо при прогнозуванні churn. Для її подолання застосовують ресемплінг (SMOTE, ADASYN) та cost-sensitive learning. Метрики precision, recall, F1-score і ROC-AUC забезпечують об'єктивну оцінку моделей, на відміну від accuracy.

Предиктивне моделювання в SaaS дозволяє проактивно підвищувати утримання, залученість і конверсію користувачів. Традиційні методи, як ARIMA, ефективні для лінійних часових рядів, тоді як для складних послідовностей застосовують глибинні мережі LSTM і

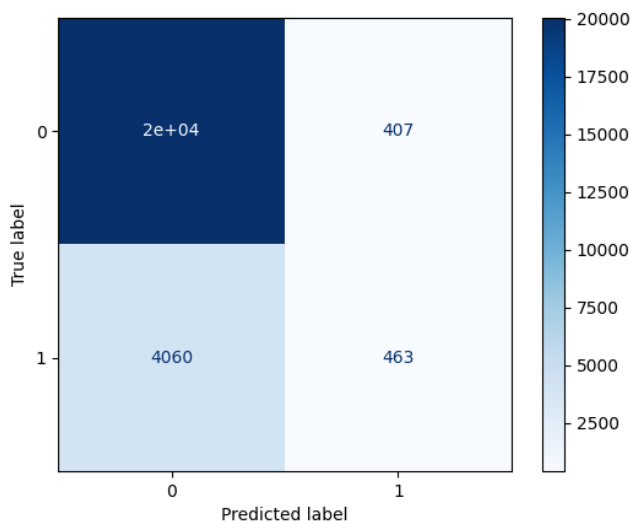
GRU, а сучасні трансформери — для точного моделювання довготривалих залежностей. Приклад — фреймворк CAWAL, що поєднує дані сесій з ансамблевими алгоритмами і досягає понад 92% точності [2].

Також застосовують аналіз виживаності (Kaplan-Meier, модель Кокса) для оцінки ризику churn і часу утримання. Методи підкріпленого навчання, MDP і HMM, допомагають моделювати динамічні користувацькі процеси та персоналізувати шляхи взаємодії. Разом ці підходи дають SaaS-платформам потужні інструменти для прогнозування і адаптації стратегій з метою сталого зростання [1].

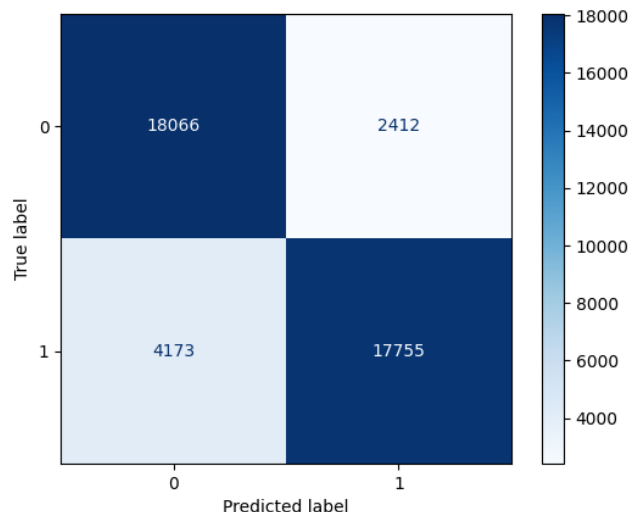
У практичному кейсі для прогнозування відтоку користувачів був використаний публічний датасет SaaS проекту і застосовані трансформерні моделі для обробки категоріальних значень для отримання багатого контексту і покращення класифікації. Також були використані AdaBoost з/та без ADASYN для балансування класів. Модель показала F1-score 0.84 після трансформації даних та балансування класів.

Табл. 1. Результати практичного дослідження

		без ADASYN			з ADASYN		
Дані	Модель	F1 train	F1 test	F1 delta	F1 train	F1 test	F1 delta
Оригінальні дані	AdaBoost	0.195	0.1882	0.0067			
Оброблені дані (Feature Scaling, Outlier Removal)	AdaBoost	0.1934	0.189	0.0044	0.1934	0.189	0.0044
Дані з конвертацією категорійних значень у вектори	AdaBoost	0.2025	0.1956	0.0069	0.8445	0.8436	0.0009



Мал. 1 Матриця невідповідностей для моделі AdaBoost на сирих даних



Мал. 2 Матриця невідповідностей для моделі AdaBoost на даних з трансформованими категорійними значеннями

Попри успіхи машинного навчання у аналізі поведінки користувачів SaaS, існують виклики: забезпечення конфіденційності та відповідність GDPR, боротьба з model drift, що знижує точність прогнозів, і складність інтерпретації моделей. Подальші дослідження мають зосередитись на причинно-наслідковому аналізі, real-time аналітиці та кросплатформних

моделях для глибокого розуміння user journeys. Загалом, машинне навчання допомагає прогнозувати поведінку користувачів, протидіяти churn і оптимізувати залучення, а майбутній фокус — на етичному AI та прозорості.

Бібліографія

1. Chakraborty, A., Raturi, V., & Harsola, S. (2022). BBE-LSWCM: A Bootstrapped Ensemble of Long and Short Window Clickstream Models. arXiv preprint arXiv:2203.16155. <https://arxiv.org/abs/2203.16155>
2. Canay, O., & Kocabicak, U. (2025). Predictive modeling and anomaly detection in large-scale web portals through the CAWAL framework. arXiv preprint arXiv:2502.00413. <https://arxiv.org/abs/2502.00413>
3. Yuan, J., Qiu, X., Wu, J., Guo, J., Li, W., & Wang, Y.-G. (2024). Integrating behavior analysis with machine learning to predict online learning performance: A scientometric review and empirical study. arXiv preprint arXiv:2406.11847. <https://arxiv.org/abs/2502.00413>
4. Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2019). BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. arXiv preprint arXiv:2001.00621. <https://arxiv.org/abs/2001.00621>