Електронний журнал «Ефективна економіка» включено до переліку наукових фахових видань України з питань економіки (Категорія «Б», Наказ Міністерства освіти і науки України № 975 від 11.07.2019). Спеціальності – 051, 071, 072, 073, 075, 076, 292.

Ефективна економіка. 2024. № 4.

DOI: http://doi.org/10.32702/2307-2105.2024.4.30
УДК 338.1

О. І. Ляшенко,
д. е. н, професор, завідувачка кафедри економічної кібернетики, Київський національний університет імені Тараса Шевченка
ORCID ID: https://orcid.org/0000-0002-0197-4179
Т. В. Кравець,
к. ф.-м. н., доцент, доцент кафедри економічної кібернетики,
Київський національний університет імені Тараса Шевченка
ORCID ID: https://orcid.org/0000-0003-4823-5143
Р. Р. Могильна,
бакалавр економіки,
Київський національний університет імені Тараса Шевченка
ORCID ID: https://orcid.org/0009-0008-7788-367X

КЛАСИФІКАЦІЯ КЛІЄНТІВ НА РАННІХ ЕТАПАХ ВИКОРИСТАННЯ МОБІЛЬНОГО ДОДАТКУ

O. Liashenko,
Doctor of Economic Sciences, Professor, Head of the Department of Economic Cybernetics, Taras Shevchenko National University of Kyiv
T. Kravets,
PhD in Physico-Mathematical Sciences, Associate Professor, Associate Professor of the Department of Economic Cybernetics, Taras Shevchenko National University of Kyiv
R. Mohylna,
Bachelor of Economics, Taras Shevchenko National University of Kyiv

CLASSIFICATION OF CUSTOMERS IN THE EARLY STAGES OF USING A MOBILE APPLICATION

У статті досліджено проблему виявлення високодохідних користувачів (китів) на ранніх етапах використання мобільного додатку, що є важливою складовою стратегії бізнесу. Розвиток продукту з трансакційною моделлю монетизації спирається на процес ідентифікації найцінніших, високодохідних та лояльних клієнтів, які стануть джерелом значної частки прибутку для компанії.

Для визначення критерію розподілу користувачів на класи була вибрана метрика, що влучно характеризує тип користувача. Нею стала виручка за перші 3 місяці використання додатку. Вона є порівнюваною для клієнтів з різних когорт та добре описує їх фінансову цінність для бізнесу. Пороговим значенням метрики для розділення користувачів на класи було обрано 150 доларів шляхом наближення до правила Парето, коли 80% виручки генерується 20% клієнтів.

Було проведено огляд структури даних та їх підготовка до моделювання методами машинного навчання для вирішення задачі класифікації. Вона включає в себе обробку відсутніх значень, викидів, видалення певних колонок, нормалізація числових даних, кодування категоріальних стовпців з допомогою OneHotEncoder, розділення на тренувальні і тестову вибірки.

Наступним кроком було навчання моделей з використанням python бібліотек sklearn та keras. Для даного дослідження обрано моделі: k найближчих сусідів (k-NN); логістична регресія; Random Forest; Support Vector Machine; нейронна мережа; CatBoost. Кожна з них має свої особливості, переваги та недоліки, можливості оптимізації шляхом підбору оптимальних гіперпараметрів.

Після порівняння різних моделей машинного навчання оптимальною моделлю виявилась CatBoost, яка правильно класифікувала 87% об'єктів. AUC (PRC) на рівні 0.78 та F1-Score = 0.66 говорять про хорошу здатність класифікувати позитивні класи та достатній баланс між точністю і повнотою. До найбільш вагомих незалежних показників моделі увійшли виручка перших трьох покупок, виручка за перші 3 дні, тривалість чатів за перші 7 днів, величина першої покупки, кількість спеціалістів, з якими спілкувався користувач протягом перших двох тижнів.

Отже, виявлення китів на ранніх етапах використання мобільного додатку є важливим для отримання високих фінансових показників компанії. Отримані результати можуть послужити основою для розвитку стратегій в сфері маркетингу, продуктового управління та підтримки клієнтів українських IT компаній.

The article examines the problem of identifying high-income users (whales) in the early stages of using a mobile application, which is an important component of a business strategy. Product development with a transactional monetization model is based on the process of identifying the most valuable, highly profitable, and loyal customers who will become the source of a significant share of profit for the company.

To determine the criterion for dividing users into classes, a metric that accurately characterizes the type of user was chosen. It was the revenue for the first 3 months of using the application. It is comparable for customers from different cohorts and well describes their financial value to the business. The threshold value of the metric for dividing users into classes was chosen to be $150 by approximating the Pareto rule, when 80% of revenue is generated by 20% of customers.

An overview of the data structure and its preparation for modeling by machine learning methods to solve the classification problem was carried out. It includes processing of missing values, and outliers, removal of certain columns, normalization of numerical data, coding of categorical columns using OneHotEncoder, and separation into training and test samples.

The next step was to train the models using the sklearn and keras python libraries. For this study, the following models were chosen: k nearest neighbors (k-NN); logistic regression; Random Forest; Support Vector Machine; neural network; and CatBoost. Each of them has its features, advantages and disadvantages, and optimization possibilities by selecting optimal hyperparameters.

After comparing different machine learning models, CatBoost proved to be the optimal model, which correctly classified 87% of objects. AUC (PRC) at the level of 0.78 and F1-Score = 0.66 indicate a good ability to classify positive classes and a sufficient balance between accuracy and completeness. The most significant

independent indicators of the model include revenue from the first three purchases, revenue for the first 3 days, duration of chats for the first 7 days, the size of the first purchase, and the number of specialists with whom the user communicated during the first two weeks.

Therefore, detecting whales in the early stages of using a mobile application is important for obtaining high financial indicators for the company. The obtained results can serve as a basis for the development of strategies in the field of marketing, product management, and customer support for Ukrainian IT companies.

Ключові слова: *трансакційна модель монетизації, кити в бізнесі, класифікація, методи машинного навчання.*

Keywords: *transactional monetization model, whales in business, classification, machine learning methods.*

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями. Ніша розробки мобільних додатків стає все більш популярною і привабливою для ведення бізнесу. Вона має значний об'єм потенційних клієнтів та є перспективною з погляду отримання високого рівня доходу. Для збереження конкурентоспроможності потрібно мати глибоке розуміння того, які саме клієнти найбільше впливають на фінансові показники, та активно працювати над їх залученням, покращенням досвіду користування та підвищенням задоволеності. Традиційно високодохідних користувачів, які генерують значну частку виторгу компанії, називають китами.

Виявлення китів на ранніх етапах використання мобільного додатку дозволяє забезпечити максимальну ефективність роботи всіх відділів та спрямувати ресурси на залучення та утримання саме цільової аудиторії, яка має великий потенціал для генерації виторгу. Це може включати розробку персоналізованих стратегій маркетингу та продажу, персоналізованих пропозицій, програм лояльності. Зосередження зусиль на високодохідних користувачах сприяє збільшенню середнього чека, частоти покупок та загального виторгу компанії. Виявлення китів на ранніх етапах дозволяє

виділити ознаки, які можуть вказувати на зростання ризику втрати цих користувачів. Це надає можливість вжити необхідних заходів для збереження їхньої лояльності та задоволеності, що сприяє збереженню прибутковості та фінансової стабільності бізнесу.

Аналіз останніх досліджень і публікацій. Моделі монетизації бізнесу включають різні стратегії заробітку коштів на продуктах або послугах. Дві загально поширені моделі монетизації – модель підписки, що надає стабільний, легко прогнозований дохід та трансакційна, що не має обмежень щодо кількості виторгу з користувача і є менш стабільною. Особливістю трансакційної моделі є наявність китів, виявлення та утримання яких має велике значення для бізнесу [1, 2, 3].

Перевагами трансакційної моделі є швидке отримання прибутку, масштабованість та відсутність обмежень щодо фізичної можливості здійснення великої кількості покупок за обмежений проміжок часу. Дана модель приваблива для компаній, оскільки дає можливість потенційно заробляти більше коштів з кожної окремої трансакції й відповідно з кожного клієнта. До недоліків трансакційної моделі монетизації відносяться нестабільність прибутку через залежність від обсягів продажу товарів чи надання послуг, залежність від постійного привернення нових клієнтів та збереження наявних. Вони спричинюють ще один недолік трансакційної моделі у вигляді складності прогнозування [2].

У роботі [4] автори розглядали механізм генерування доходу з метою забезпечення фінансової стабільності для розвитку бізнесу. Було розроблено загальну концепцію моделі доходу та на основі множинного лінійного регресійного аналізу визначено змінні, які сильно впливають на дохід.

Дослідження споживчих звичок користувачів мобільних додатків серед представників різних поколінь було проведено авторами у роботі [5]. На основі статистичного аналізу було визначено поведінкові характеристики цільових поколінь, розглянуто особливості мобільного маркетингу та мобільних додатків.

Дана робота застосовує методи машинного навчання для моделювання показника, який дозволить виділити китів серед загальної групи користувачів

мобільного додатку.

Формулювання цілей статті (постановка завдання). Метою роботи є визначення показників, що найбільше характеризують потенційно високодохідних користувачів на ранніх етапах користування мобільним додатком, та проведення класифікації користувачів методами машинного навчання.

Виклад основного матеріалу дослідження. Дані, використані в цьому дослідженні надані українською IT компанією, флагманським продуктом якої є мобільний додаток що надає послуги онлайн спілкування з експертами [6].

Перш за все потрібно визначити хто такий кит за допомогою числової характеристики. Для задачі класифікації користувачів трансакційної моделі, можна використовувати такі ознаки як загальна кількість трансакцій, сума грошових операцій, частота використання сервісу, активність в певному періоді часу тощо. За допомогою певних порогових значень цих ознак можна виділити, китів із загальної групи користувачів. Варто також зауважити що визначення ознак значно залежить від самого продукту та ключових послуг, які він надає.

У процесі вибору цільової метрики для даного дослідження були розглянуті наступні показники, які можуть свідчити про активність та значущість користувача в додатку: кількість часу проведеного в додатку; кількість покупок; кількість чатів зі спеціалістами; виторг, який приніс користувач.

З урахуванням можливих спекуляцій та обґрунтованості, вибір для цього дослідження був зроблений на користь метрики виторгу. Ця метрика є простою, зрозумілою та точно відображає внесок користувача в бізнес. Її використання допомагає встановити чіткі цілі та спрямувати зусилля на досягнення фінансових показників у контексті розуміння користувачів та їх впливу на продукт.

Варто взяти до уваги, що кожен користувач прийшов в додаток у різний період часу, тому вони знаходяться в різних умовах. Наприклад, користувач, який прийшов рік тому, мав цілий рік на те, щоб користуватись додатком і генерувати виторг для компанії, на відміну від користувача, який прийшов

місяць тому. Отже, для того, щоб створити однакові умови для всіх клієнтів, було обрано вікно в 3 місяці, протягом яких рахувався виторг з користувача.

З аналізу залежності між кумулятивним виторгом і кумулятивною кількістю користувачів, можна зробити наступні висновки. Користувачі, у яких величина виторгу перевищує 150 доларів, складають лише 26% від усіх клієнтів, проте вони приносять 82% від загального доходу компанії. Це свідчить про велику цінність цієї групи користувачів для бізнесу. Тому можна зробити висновок, що виторг понад 150 доларів є важливим показником, який відокремлює високодохідних користувачів від інших. Оскільки цей поріг виторгу є значущим з погляду впливу на загальний дохід компанії, його можна використовувати для трансформації грошової змінної на бінарну. Така бінарна змінна буде слугувати залежною змінною для подальшого моделювання та класифікації користувачів на китів і не китів на ранніх етапах життя продукту.

Набір даних складається з 34 колонки, 8 з них типу object та 26 float/int. Розглянемо ці характеристики користувача з розподілом по типу.

Технічні характеристики:

- user_id – унікальний ідентифікатор користувача;
- media_source – маркетинговий канал з якого був залучений користувач;
- platform – платформа користувача (iOS, Android);
- pushes_enabled – бінарна величина, що відповідає на запитання, чи дав користувач доступ на відправку йому push-повідомлень;
- language – мова, яка використовується в додатку;
- paying – бінарна величина, що відповідає на питання, чи є користувач платником в інших сервісах в додатку, що не стосуються досліджуваних онлайн консультацій.

Демографічні характеристики:

- country – країна користувача;
- age – вік;
- gender – стать;
- relationship – статус відносин, в якому знаходиться користувач.

Поведінкові характеристики.

Тут знаходяться дані про поведінку користувача протягом перших 2 тижнів його життя в додатку. Це було зроблено для того, щоб всіх користувачів поставити в однакові умови, а також, якщо в майбутньому ці показники будуть слугувати вхідними даними для моделі, то для їх збору потрібно було відносно небагато часу. Серед них теж можна виділити підгрупи: характеристики що стосуються поповнень користувача; та ті, які стосуються чатів користувача.

- revenue_3m – виторг, який приніс користувач за перші три місяці (надалі буде використана як база для створення залежної змінної моделей);
- first_chat_day – кількість днів що пройшла він встановлення додатка до першого чату;
- first_chat_type – тип першого чата (онлайн / офлайн / месенджер);
- first_chat_duration – довжина першого чата в секундах;
- first_3chats_duration – довжина перших трьох чатів в секундах;
- chats_duration_7d – довжина чатів протягом перших 7 днів в додатку;
- chats_duration_2w – загальна тривалість чатів за перші 2 тижні після встановлення додатку;
- chats_number_2w – кількість чатів користувача за перші 2 тижні після встановлення додатка;
- specialists_count_2w – кількість спеціалістів, з якими спілкувався користувач протягом перших 2 тижнів;
- avg_review_rate_2w – середня оцінка користувачем його чатів зі спеціалістами протягом перших 2 тижнів;
- first_purchase_day – кількість днів що пройшла від встановлення додатка до першої покупки;
- first_purchase_revenue – виторг з першої покупки користувача;
- third_purchase_day – кількість днів, що пройшла з встановлення додатка до третьої покупки;
- first_3purchases_revenue – виторг з перших трьох покупок;
- revenue_1d – виторг, який приніс користувач у свій перший день;

- revenue_3d – виторг, який приніс користувач в перші три дні;
- revenue_7d – виторг за перші 7 днів;
- revenue_2w – виторг за перші 2 тижні;
- purchases_number_2w – кількість покупок користувача в перші 2 тижні;
- paypal_purchases_2w – чи були в користувача покупки методом PayPal;
- google_pay_purchases_2w – чи були покупки методом Google Pay;
- card_purchases_2w – чи були покупки які оплачувались картою;
- recurrent_purchases_2w – чи були в користувача рекурентні платежі (повторні платежі певним видом оплати);
- applepay_purchases_2w – чи були покупки методом Apple Pay.

Дані були зібрані та опрацьовані через базу даних Vertica. *Vertica Database* — це стовпчаста база даних, призначена для обробки великих обсягів даних і забезпечення швидкої роботи [7].

Розглянемо детальніше структуру даних. Найбільший відсоток китів серед усіх користувачів спостерігається для країн Індонезія, Аргентина, Пакистан. Проте в абсолютних величинах найбільша кількість перспективних клієнтів припадає на США та Канаду. Якщо зосередитися на відсотку топ платників з розбивкою по джерелу залучення трафіку, то серед користувачів, що прийшли органічно (самостійно шукали продукт), відсоток високоприбуткових користувачів вищий ніж серед тих, хто залучений на платній основі. Виявлена значно більше китів як у відсотковій, так і в абсолютній кількості на платформі iOS. Також можна помітити, що серед користувачів, які не дозволили відправляти собі push-повідомлення, більший відсоток китів. Проте абсолютне значення вище для групи, яка дала дозвіл на push-повідомлення. Наступною ознакою є мова. Найбільший відсоток китів виявлено для іспаномовної аудиторії, причому власне користувачів у цій групі небагато. Розподіл за статтю дав практично однаковий відсоток китів, але в абсолютної кількості користувачів більше жінок.

Продукт, користувачі якого досліджуються, є комплексним і містить не лише консультації, а й інші платні послуги. Виявлено, що саме серед платників відсоток китів більший, 32% проти 23% для не платників. Поясненням цьому є те, що користувачі, які вже заплатили, мають більшу зацікавленість продуктом та є більш платоспроможними. Якщо дивитись на тип першого чату, то найбільш ефективним з погляду активації є онлайн чат, коли спеціаліст і клієнт спілкуються в реальному часі. Сімейний статус говорить про те, що більш перспективними є групи без пари та ті, хто вважає, що в нього складнощі у відносинах. Впливу вікової характеристики не помітно.

Виявлено значну різницю в тривалості першого чату, перших трьох чатів, тривалості чатів за перші 7 та 14 днів після встановлення додатка для групи китів та звичайних користувачів. Те ж саме можна сказати й про кількість чатів за перші два тижні. Показник кількості спеціалістів є вищим для групи китів, але є високодохідні користувачі, які мали чат лише з одним спеціалістом.

Перейдемо до показників, пов’язаних з виторгом. Ці показники досить сильно корелюють між собою, тому надалі потрібно буде видалити частину з них.

Показник кількості здійснених покупок суттєво відрізняється для високодохідних і звичайних користувачів. Кількість покупок має більше медіанне значення та більший розмах для китів. Закономірно, що для користувачів, які мали хоча б одну покупку за типом оплати PayPal, card, Apple Pay, Google Pay, імовірність стати китом буде вищою. Цікавим показником є наявність рекурентних платежів. Для нього різниця відсотків китів суттєво відрізняється: 57% китів у групі, що мала повторні платежі та 24% китів у групі, що не мала. Отже, наявність рекурентних покупок може бути інформативним показником для моделі, як і показники виторгу та тривалості чатів.

На цьому етапі було проведено попередню обробку даних, їх очищення від викидів, заповнення або видалення рядків/стовпців з відсутніми значеннями [8]. В ході підготовки набору даних були видалені наступні нерелевантні колонки:

- user_id – унікальний ідентифікатор користувача, який не дає жодної інформації про його характеристики чи поведінку;
- first_chat_day – має близько половини значень null.
- first_purchase_day – як і first_chat_day видалено через велику кількість пропущених значень.

Деякі категоріальні змінні були згруповані для зручності подальшої стандартизації:

- relationship – complicated, in_relationship, single
- country - United_States, Canada, Other, Australia, United_Kingdom
- media_source – всі значення колонки були віднесені до одної з груп: Organic, Facebook, Apple Search Ads, Others, Google, TikTok.

Заповнення пропущених значень:

- media_source заповнені значенням ‘Unknown’ оскільки це користувачі з не відстеженим джерелом, що може виникати з певних причин, наприклад заборона на збір даних;
- country, language, gender, relationship – порожні значення були заповнені модою;
- pushes_enabled – порожні значення заповнені медіаною.

Для подальшого моделювання було створено нову залежну змінну revenue_3m_binary на основі revenue_3m з розбиттям на рівні 150 доларів. Було виявлено, що розподіл залежної змінної є не рівномірним оскільки китів значно менше ніж звичайних користувачів. Цього дисбалансу можна позбутися різними шляхами, наприклад методом повторної вибірки штучно збільшити кількість користувачів у класі китів. Також можна випадково відрізати частину більшого класу. Якщо використовувати другий варіант, то даних лишиться мало для подальшої роботи з ними. Якщо ж збільшувати менший клас, то через великий дисбаланс це може призвести до викривлення показників оцінок моделі оскільки один і той же приклад буде зустрічатись багато разів [9]. Отже, було прийнято рішення не збалансовувати модель і враховувати це при оцінці моделі.

Після дослідження на мультиколінеарність було видалено показники: first_chat_duration; chats_duration_2w; revenue_7d; revenue_1d; chats_number_2w, purchases_number_2w, revenue_2w.

Наступні колонки мають лише по два значення, тому були закодовані в бінарні цілі числа: platform, pushes_enabled.

Категоріальні дані, що приймають понад два значення, розділені на нові змінні, кожна з яких відповідає окремому значенню початкової колонки: language, gender, first_chat_type, media_source_group, country_group, relationship_group. Для цього було використано метод OneHotEncoder [10]. Однак, при використанні цього методу, існує ризик виникнення проблеми мультиколінеарності, що може стати проблемою при використанні деяких моделей машинного навчання, таких як лінійна регресія. Для уникнення цієї проблеми було видалено одну з бінарних змінних.

До числових даних було застосовано Min-Max Scaling, тобто переведені в діапазон від 0 до 1. Останнім підготовчим етапом є розділення на тестову і тренувальну вибірки за допомогою функції. Для тестування моделей було залишено 30% вибірки.

На наступному етапі були побудовані моделі класифікації користувачів трансакційної моделі на китів та звичайних користувачів на ранніх етапах життя продукту. Розглянуто застосування шести моделей машинного навчання та порівняно їх показники оцінки адекватності та якості класифікації [11].

К найближчих сусідів (k-NN) [12]. Результати застосування моделі показують прийнятну точність на рівні 0.79. Значення AUC (ROC) для даної моделі становить 0.75, що свідчить про відносно хорошу здатність розрізняти класи. Проте згадаємо про те, що маємо справу з розбалансованими класами, тому звернемо увагу на криву Precision-Recall, що використовується для вимірювання точності та повноти. Значення AUC (PRC) дорівнює 0.59, що може вказувати на меншу здатність моделі до визначення позитивних екземплярів. F1 Score, який говорить про співвідношення точності та повноти, має значення 0.51, що є досить низьким показником. Враховуючи ці результати, модель k-NN можна спробувати покращити за допомогою налаштування гіперпараметра k.

Для оптимізації алгоритму було використано GridSearchCV - метод визначення оптимальних гіперпараметрів моделі шляхом перебору всіх можливих комбінацій заздалегідь заданого простору [13]. Використовуючи крос-валідацію він знаходить комбінацію що дасть найкращу точність моделі. Для покращення даного алгоритму були перебрані кількості найближчих сусідів у межах від 3 до 30 з кроком 1. В результаті отримали оптимальне значення параметра 17. Оптимізована модель k-NN досягає точності 0.81, AUC (ROC) 0.78. Проте якщо дивитись на показник AUC (PRC), то він покращився всього лише до рівня 0.62, а F1-Score взагалі не змінився, що говорить про недостатню точність моделі.

Логістична регресія правильно класифікує 83% прикладів. При цьому AUC рівний 0.76, що вказує на те, що модель має добру роздільну здатність. Значення AUC (PRC) рівне 0.66 та вказує на помірну здатність моделі зберігати точність при визначенні позитивних класів. F1-Score = 0.60, що означає помірний баланс між точністю та повнотою моделі. Проте якщо порівнювати з попереднім алгоритмом, то значення гармонічного середнього вже є значно кращим, беручи до уваги також те, що позитивний клас є значно меншим і його досить важко класифікувати.

Перевагою логістичної регресії є її простота в інтерпретації, тож подивимось, яким ознакам вона надала найбільшу вагу. На фінальний результат найбільше впливають виторг перших трьох покупок, тривалість чатів за перші 7 днів та виторг за три дні. Також вагомими ознаками є: кількість спеціалістів, з якими користувач мав чати; наявність третьої покупки; вік. Це свідчить про те, що високодохідність користувача проявляється вже на першому тижні перебування на продукті.

Спробуємо оптимізувати модель шляхом підбору гіперпараметрів. Для цього знову використаємо GridSearchCV та спробуємо підібрати потрібні penalty, що будуть штрафувати модель за дуже низькі або високі коефіцієнти перед змінними та параметр C, який є оберненою силою регуляризації. Оптимальним виявилось використовувати регулятор L1 з C = 1.9. Це дало наступні результати моделі. Загальна точність не покращилась, залишившись на рівні 0.83, проте AUC (ROC) зріс до 0.77. При цьому AUC (PRC) покращився

до 0.68, а F1 Score навпаки зменшився з 0.60 до 0.58 після підбору гіперпараметрів. Оскільки ми маємо розбалансовані дані, то показник F1 є досить важливим, що не дає змоги використовувати дану модель для класифікації.

Модель *Random Forest (RF)* показала високу точність на рівні 0.84 та здатність розділити класи, про що свідчить AUC (ROC) на рівні 0.82. Також модель має добру здатністю зберігати точність при виявленні позитивних класів (AUC (PRC) = 0.75). F1-Score 0.65 вказує на відносно хороший баланс між точністю та повнотою моделі. Враховуючи ці показники, можна стверджувати, що модель RF показує хорошу ефективність при класифікації.

Для того, щоб оптимізувати RF переберемо гіперпараметри, які використовуються в моделі за допомогою RandomizedSearchCV, який обирає випадкові набори параметрів і шукає найкращі [14]. В результаті отримано такий оптимальний набір параметрів: 590 дерев, кількість ознак що рівна квадратному кореню від загальної кількості ознак, глибина дерев на рівні 10, метрика оптимальності розбиття entropy та використання бутстрапінгу. Показники оптимізованої моделі: точність моделі 0.87; площа під кривою ROC 0.86, що є досить високим показником; AUC(PRC) покращилось до 0.78; F1-Score 0.65 лишився на тому ж рівні.

Support Vector Machine [15] показав добру точність на рівні 0.82 та здатність розділити класи AUC (ROC) 0.79, але мав помірну здатність зберігати точність при виявленні позитивних класів (AUC (PRC) = 0.67). F1-Score виявився рівним 0.59. Ці показники є досить низькими в порівнянні з іншими моделями, тому ми не будемо використовувати SVM для класифікації китів та звичайних користувачів.

Перейдемо до моделі *Нейронної мережі (Sequential model)* [16]. Загалом, виходячи з матриці невідповідностей вона показала високу точність і правильно класифікувала 83% прикладів. Якщо говорити про її здатність розділити класи, то вона мала AUC (ROC), рівний 0.84. Разом з цим вона продемонструвала добру здатність зберігати точність при виявленні позитивних класів (AUC (PRC) = 0.73). F1-Score, рівний 0.63, також вказує на помірний баланс між точністю та повнотою моделі. Незважаючи на досить

гарні показники, вони є гіршими ніж, наприклад у моделі RF. Імовірніше за все нейронна мережа показала гірші результати через недостатньо велику кількість даних.

Наступна модель – *CatBoost* [17]. Якщо говорити про загальну точність, то вона правильно класифікувала 86% прикладів, що є досить високим показником. При цьому вона мала AUC = 0.85, AUC (PRC) = 0.77. F1-Score, що рівний 0.66, також вказує на хороший баланс між точністю та повнотою моделі.

Спробуємо покращити точність моделі завдяки налаштуванням гіперпараметрів. Перший з них це швидкість навчання. Встановлення меншого значення може допомогти покращити збіжність моделі, але може зайняти більше часу для тренування.

Для підбору використаємо значення 0.01, 0.1 та 0.5, що відповідають повільному, середньому та швидкому навчанню. Також переберемо глибину дерев в діапазоні від 4 до 9. Збільшення глибини дерева може дозволити моделі вчитися більш складним залежностям у даних, але при цьому може збільшитися ймовірність перенавчання. В результаті отримали значення глибини дерев на рівні 6 та швидкості навчання 0.01, тобто повільному.

Застосування моделі з такими гіперпараметрами дає змогу правильно класифікувати 87% прикладів, що на 1% більше ніж до оптимізації. Також відбулося збільшення показника AUC (ROC) до рівня 0.87. При цьому трохи підвищилась здатність точно класифікувати позитивні приклади, оскільки AUC (PRC) рівний 0.78. F1-Score =0.66 залишився на тому ж рівні та говорить про гарний баланс між точністю і повнотою моделі.

Наведемо перелік ознак, які мали найбільший вплив на результати класифікації: виторг перших трьох покупок користувача; виторг за перші 3 дні на продукті; тривалість чатів за перші 7 днів; величина першої покупки; кількість спеціалістів, з якими спілкувався користувач протягом перших двох тижнів. Це свідчить про те, що користувач уже при перших діях своєю поведінкою показує що він є платоспроможним і готовий витрачати багато. Якщо говорити про демографічні показники, то вони знаходяться значно нижче по списку та мають менший вплив на результат моделі. Це говорить про те, що ці ознаки значно менше визначають імовірність стати китом. Проте вони все ж

входять в топ 15, зокрема вік користувача, країна, статус відносин, джерело залучення.

Для вибору оптимальної моделі класифікації користувачів на китів та не китів порівняємо показники якості для кожної з побудованих моделей (Табл. 1). Бачимо, що за всіма ціновими показниками й видами оцінок, найкращі результати показав алгоритм CatBoost та CatBoost (Оптимізований).

Таблиця 1. Порівняння показників якості моделей

Модель	Accuracy	AUC (ROC)	AUC (PRC)	F1 Score
k-NN	0.79	0.75	0.59	0.51
k-NN (Оптимізована)	0.81	0.78	0.62	0.51
Логістична регресія	0.83	0.76	0.66	0.60
Логістична регресія (Оптимізована)	0.83	0.77	0.68	0.58
Random Forest	0.84	0.82	0.75	0.65
Random Forest (Оптимізована)	0.84	0.85	0.76	0.64
Support Vector Machine	0.82	0.79	0.67	0.59
Нейронна мережа	0.83	0.84	0.73	0.63
CatBoost	0.86	0.85	0.77	0.66
CatBoost (Оптимізована)	0.87	0.87	0.78	0.66

Джерело: створено авторами

Результати проведеного дослідження є важливим внеском у розуміння поведінки користувачів на ранніх етапах життя на продукті. Виявлений вплив вхідних ознак на вихідну в рамках побудованої моделі надає цінну інформацію для команди маркетингу, що займається залученням користувачів, продуктовій команді з погляду визначення стратегій активації та утримання та службі підтримки користувачів.

В цілому, розуміння впливу окремих ознак на модель може допомогти в розробці стратегій та плануванні дій в різних аспектах бізнесу. Це дозволяє зосередитися на ключових факторах, що впливають на важливі метрики успіху та підвищують ефективність маркетингових та продуктових дій.

Комбінування цих результатів з іншими дослідженнями та даними може сприяти покращенню ефективності бізнес-процесів та досягненню більшого задоволення клієнтів.

Висновки та перспективи подальших розвідок у даному напрямі.

Визначення китів на ранніх етапах використання мобільного додатку є важливою складовою стратегії бізнесу та розвитку продукту з трансакційною моделлю монетизації. Це процес ідентифікації найцінніших, високодохідних та лояльних клієнтів, які стануть джерелом значної частки прибутку для компанії. Правильне виявлення китів та сфокусована робота з такими користувачами на самому початку їхнього досвіду користування продуктом є запорукою успішності бізнесу та рентабельності проєкту загалом.

Для визначення критерію розподілу користувачів на класи була вибрана метрика, що влучно описує китовість користувача. Нею став виторг за перші 3 місяці використання додатка. Вона є порівнюваною для клієнтів з різних когорт та добре описує фінансову цінність для бізнесу. Пороговим значенням для розділення користувачів на китів та не китів було обрано 150 доларів шляхом наближення до правила Парето, коли 80% виторгу генерується 20% клієнтів.

Було проведено огляд структури даних та їх підготовка до моделювання методами машинного навчання для розв’язання задачі класифікації. Для даного дослідження обрано такі моделі як k найближчих сусідів, логістична регресія, випадковий ліс, Support Vector Machine, нейронна мережа, CatBoost. Кожна з них має свої особливості, переваги та недоліки, можливості оптимізації шляхом підбору оптимальних гіперпараметрів.

Після порівняння різних моделей машинного навчання оптимальною моделлю виявилась CatBoost, яка правильно класифікувала 87% об’єктів. AUC (PRC) на рівні 0.78 та F1-Score = 0.66 говорять про хорошу здатність класифікувати позитивні класи та достатній баланс між точністю і повнотою. До найбільш вагомих незалежних показників моделі увійшли виторг перших трьох покупок, виторг за перші 3 дні, тривалість чатів за перші 7 днів, величина першої покупки, кількість спеціалістів, з якими спілкувався користувач протягом перших двох тижнів.

Отже, виявлення китів на ранніх етапах використання мобільного додатку є важливим для отримання високих фінансових показників компанії. Отримані результати можуть послужити основою для розвитку стратегій у сфері маркетингу, продуктового управління та підтримки клієнтів українських IT

компаній. Надалі планується розглянути спроможність розглянутих та інших методів машинного навчання класифікувати користувачів з різних баз даних.

Література

1. Pereira, D. (2022), *Revenue Models*, The Business Model Analyst, available at: https://businessmodelanalyst.com/wp-content/uploads/woocommerce_uploads/2021/03/Revenue-Models-kjr2dm.pdf (Accessed 24 Nov 2023).

2. Capgemini (2017), "Predictive Modeling Using Transactional Data", available at: https://www.capgemini.com/wp-content/uploads/2017/07/Predictive_Modeling_Using_Transactional_Data.pdf (Accessed 15 Nov 2023).

3. Liverence, B. (2017), "Whale Watching: Many companies earn a huge portion of sales from a few customers", *Bloomberg Second Measure*, available at: https://secondmeasure.com/datapoints/whales/ (Accessed 12 Nov 2023).

4. Remenová, K. Kintler, J. and Jankelová, N. (2020), "The General Concept of the Revenue Model for Sustainability Growth", *Sustainability*, vol. 12, 6635, available at: doi:10.3390/su12166635 (Accessed 4 Dec 2023).

5. Bencsik, A., Machová, R. and Zsigmond, T. (2018), "Analysing customer behaviour in mobile app usage among the representatives of Generation X and Generation Y", *Journal of Applied Economic Sciences*, vol. XIII, Fall 6(60), pp. 1668-1677.

6. Products. Obrio, available at: https://obrio.co/products (Accessed 20 Jan 2024).

7. Vertica (2024), "Vertica. Powering the World's Data-Driven Leaders", available at: https://www.vertica.com/about/ (Accessed 18 Jan 2024).

8. Joseph, A. Nelson, B. Rubinstein, B. Tygar, J. (2019), *Adversarial Machine Learning*, Cambridge University Press, doi: 10.1017/9781107338548.

9. Liashenko, O. Kravets, T. and Kostovetskyi, Y. (2023) "Machine Learning and Data Balancing Methods for Bankruptcy Prediction", *Ekonomika*, vol. 102(2), pp. 28–46. doi: 10.15388/.

10. Pedregosa, F. et al. (2011), "Scikit-learn: Machine Learning in Python", *JMLR*, vol. 12, pp. 2825-2830.

11. Mello, R. Ponti, M. (2018), *Machine Learning: A Practical Approach on the Statistical Learning Theory*, Springer.

12. Zhang, S. Li, X. Zong, M. Zhu, X. and Wang, R. (2018), "Efficient kNN

Classification with Different Numbers of Nearest Neighbors" in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774-1785, doi: 10.1109/TNNLS.2017.2673241.

13. Cherif, W. (2018), "Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis", *Procedia Computer Science*, vol. 127, pp. 293-299.

14. Ramadhan, M. Sitanggang, I. Nasution, F. and Ghifari, A. (2017), "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency", *DEStech Transactions on Computer Science and Engineering*, doi: 10.12783/dtcse/cece2017/14611.

15. Guyon, I. Weston, J. Barnhill, S. et al. (2002), "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, vol. 46, pp. 389–422, doi: 10.1023/A:1012487302797.

16. Kriegeskorte, N. Golan, T. (2019), "Neural network models and deep learning", *Current Biology*, vol. 29, no. 7, pp. R231–R236. doi: 10.1016/j.cub.2019.02.034.

17. Kononenko, V. Krasnoshlyk, N. (2022), "Using boosting methods for machine learning problems", *Cherkasy university bulletin: applied mathematics. Informatics*, no. 1, pp. 58–68. doi: 10.31651/2076-5886-2021-1-58-68.

References

1. Pereira, D. (2022), Revenue Models, The Business Model Analyst, available at: https://businessmodelanalyst.com/wp-content/uploads/woocommerce_uploads/2021/03/Revenue-Models-kjr2dm.pdf (Accessed 24 Nov 2023).

2. Capgemini (2017), "Predictive Modeling Using Transactional Data", available at: https://www.capgemini.com/wp-content/uploads/2017/07/Predictive_Modeling_Using_Transactional_Data.pdf (Accessed 15 Nov 2023).

3. Liverence, B. (2017), "Whale Watching: Many companies earn a huge portion of sales from a few customers", *Bloomberg Second Measure*, [Online], available at: https://secondmeasure.com/datapoints/whales/ (Accessed 12 Nov 2023).

4. Remenová, K. Kintler, J. and Jankelová, N. (2020), "The General Concept of the Revenue Model for Sustainability Growth", *Sustainability*, [Online], vol. 12, 6635, available at: doi:10.3390/su12166635 (Accessed 4 Dec 2023).

5. Bencsik, A., Machová, R. and Zsigmond, T. (2018), "Analysing customer

behaviour in mobile app usage among the representatives of Generation X and Generation Y", *Journal of Applied Economic Sciences*, vol. XIII, Fall 6(60), pp. 1668-1677.

6. Obrio (2024), "Products", available at: https://obrio.co/products (Accessed 20 Jan 2024).

7. Vertica (2024), "Powering the World's Data-Driven Leaders", available at: https://www.vertica.com/about/ (Accessed 18 Jan 2024).

8. Joseph, A. Nelson, B. Rubinstein, B. and Tygar, J. (2019), *Adversarial Machine Learning,* Cambridge University Press, doi: 10.1017/9781107338548.

9. Liashenko, O. Kravets, T. and Kostovetskyi, Y. (2023) "Machine Learning and Data Balancing Methods for Bankruptcy Prediction", *Ekonomika*, vol. 102(2), pp. 28–46. doi: 10.15388/.

10. Pedregosa, F. (2011), "Scikit-learn: Machine Learning in Python", *JMLR*, vol. 12, pp. 2825-2830.

11. Mello, R. Ponti, M. (2018), *Machine Learning: A Practical Approach on the Statistical Learning Theory*, Springer.

12. Zhang, S. Li, X. Zong, M. Zhu, X. and Wang, R. (2018), "Efficient kNN Classification with Different Numbers of Nearest Neighbors" in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774-1785, doi: 10.1109/TNNLS.2017.2673241.

13. Cherif, W. (2018), "Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis", *Procedia Computer Science*, vol. 127, pp. 293-299.

14. Ramadhan, M. Sitanggang, I. Nasution, F. and Ghifari, A. (2017), "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency", *DEStech Transactions on Computer Science and Engineering*, doi: 10.12783/dtcse/cece2017/14611.

15. Guyon, I. Weston, J. and Barnhill, S. (2002), "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, vol. 46, pp. 389–422, doi: 10.1023/A:1012487302797.

16. Kriegeskorte, N. and Golan, T. (2019), "Neural network models and deep learning", *Current Biology*, vol. 29, no. 7, pp. R231–R236. doi: 10.1016/j.cub.2019.02.034.

17. Kononenko, V. and Krasnoshlyk, N. (2022), "Using boosting methods for machine learning problems", *Cherkasy university bulletin: applied mathematics. Informatics*, vol. 1, pp. 58–68. doi: 10.31651/2076-5886-2021-1-58-68.

Стаття надійшла до редакції 08.04.2024 р.