

АНАЛІЗ ЗАСОБІВ АВТОМАТИЗОВАНОГО МАШИННОГО НАВЧАННЯ ДЛЯ ЗАСТОСУВАННЯ У МАРКЕТИНГУ

У статті досліджено проблему автоматизації діяльності ІТ-експертів з машинного навчання за допомогою сучасних фреймворків AutoML (AutoSklearn та TPOT). Метою роботи є подолання фундаментального протиріччя між високою ресурсомісткістю ручного створення предиктивних конвеєрів та необхідністю будувати точні моделі в умовах обмеженої кількості даних, коли компанії володіють лише базовою транзакційною інформацією. Запропонований підхід формалізує використання алгоритмів AutoML для розв'язання задачі прогнозування відтоку клієнтів у роздрібній торгівлі, замінюючи ручні процеси обробки даних автоматизованими рішеннями. Підхід реалізовано шляхом генерації набору предметно-орієнтованих ознак на базі RFM-моделі та валідовано методами історичної симуляції на транзакційному датасеті «Online Retail». Експериментальні результати демонструють, що системи AutoML здатні ефективно працювати із «сирими» даними: AutoSklearn забезпечує стабільну зважену F1-міру на рівні 0.78 та ROC AUC 0.792 вже за 5 хвилин роботи. Робота має практичне значення для розробки ресурсоефективних предиктивних систем, мінімізації впливу людського фактора та пришвидшення розгортання моделей на підприємствах із базовим рівнем збору даних.

Ключові слова: автоматизоване машинне навчання, предиктивна аналітика, прогнозування відтоку клієнтів, роздрібна торгівля, інженерія ознак.

О.В. Nikonov

ANALYSIS OF AUTOMATED MACHINE LEARNING TOOLS FOR APPLICATION IN MARKETING

The article investigates the problem of automating the activities of IT experts in machine learning using modern AutoML frameworks (AutoSklearn and TPOT). The aim of the work is to overcome the fundamental contradiction between the high resource intensity of manual creation of predictive pipelines and the need to build accurate models in conditions of limited data, when companies have only basic transactional information. The proposed approach formalizes the use of AutoML algorithms to solve the problem of predicting customer churn in retail, replacing manual data processing processes with automated solutions. The approach is implemented by generating a set of subject-oriented features based on the RFM model and validated by historical simulation methods on the transactional dataset "Online Retail". Experimental results demonstrate that AutoML systems are able to work effectively with "raw" data: AutoSklearn provides a stable weighted F1-measure at the level of 0.78 and ROC AUC 0.792 in just 5 minutes of work. The work has practical significance for developing resource-efficient predictive systems, minimizing the impact of the human factor, and accelerating the deployment of models at enterprises with a basic level of data collection.

Keywords: automated machine learning, predictive analytics, customer churn prediction, retail, feature engineering.

Вступ

Сучасний конвеєр машинного навчання (ML pipeline) є доволі складним багатостадійним ІТ-процесом. Традиційне створення предиктивних моделей вимагає глибокої експертизи в програмуванні та статистиці для відбору ознак, вибору алгоритмів і тонкого налаштування гіперпараметрів [1, 2]. На практиці ІТ-фахівці та інженери з даних витрачають від 60% до 80% свого робочого часу саме на рутинні завдання з підготовки масивів інформації, що робить розробку тривалою та дорогою [3, 4]. Вирішенням цієї технологічної проблеми є впровадження засобів автоматизованого машинного навчання (AutoML). Вони зводять до мінімуму ручне втручання, самостійно тестуючи методи очищення даних та вибираючи оптимальні архітектури моделей (наприклад, ансамблі дерев рішень чи нейронні мережі). Це дозволяє суттєво скоротити час розробки та делегувати створення систем експертам без глибоких навичок кодування [5, 6, 7].

Завдяки здатності ефективно обробляти гігантські масиви даних та автоматизувати процеси, ці ІТ-рішення активно впроваджуються у маркетинг для переходу до гіперперсоналізації та підвищення рентабельності інвестицій [8, 9, 10, 11]. Критичним стратегічним викликом у цій сфері є проблема відтоку клієнтів. Залучення нових споживачів коштує компаніям у 5-7 разів дорожче, ніж утримання існуючих, а глобальні показники відтоку у висококонкурентних галузях сягають 15-25% щорічно [12, 13]. Зниження рівня відтоку лише на 5% здатне підвищити прибутковість бізнесу на 25-125%, тоді як загальні втрати від переходу клієнтів до конкурентів вимірюються трильйонами доларів [12].

Попри загальну ефективність AutoML, стандартні універсальні системи часто зосереджуються на базовій обробці табличних даних і не здатні самостійно конструювати складні предметно-орієнтовані ознаки [1, 3, 14]. З огляду на це, ідеальним і надзвичайно складним середовищем для перевірки таких ІТ-рішень є прогнозування відтоку клієнтів у роздрібній торгівлі [15]. Використання спеціалізованих плат-

форм AutoML дозволяє автоматично генерувати специфічні маркетингові метрики: показники поведінки RFM (давність, частота, сума покупок), індикатори довічної цінності (CLV) та оцінки поведінкової залученості [1, 12, 16]. Завдяки здатності алгоритмів знаходити приховані нелінійні зв'язки, компанії можуть виявляти вразливих клієнтів із безпрецедентною точністю, формувати списки ризику для цільових компаній і ефективно замінювати ручну попередню обробку даних надійною автоматизацією.

1. Аналіз існуючих проблем

Хоча дослідження у сфері автоматизованого машинного навчання (AutoML) та прогнозування відтоку клієнтів демонструють значний науковий прогрес, існує низка суттєвих недоліків та розривів між теорією і практикою, які обмежують їхнє застосування. Однією з проблем є те, що багато досліджень не розголошують дані, на яких тренувалися їхні моделі, оскільки вони переважно є приватними, що робить практично неможливим відтворення результатів та їхнє об'єктивне порівняння між різними науковими роботами. Це помітно в [1, 2, 13], про дану проблему також зазначено в [16].

Ті ж набори даних, які знаходяться у відкритому доступі, зазвичай надаються вже у попередньо обробленому або агрегованому вигляді. Це створює додаткові перешкоди, оскільки не дозволяє дослідникам виводити специфічні та нові ознаки безпосередньо на основі необробленої інформації [16]. Крім того, такий формат даних викликає обґрунтовані питання до коректності методології тестування, оскільки багато моделей оцінюються за допомогою простого випадкового розділення даних (random split), що є некоректним для ринків, які постійно змінюються. Замість цього кращою практикою є тестування на даних поза вибіркою з урахуванням часу (out-of-sample testing або історична симуляція), яке краще відображає реальні умови прогнозування, але рідше застосовується через попередню агрегацію [15].

Іншим вагомим недоліком є галузевий дисбаланс у фокусі наукових робіт. Більшість досліджень AutoML у сфері прогнозування відтоку клієнтів віддає значну перевагу таким індустріям, як фінанси (банкінг) [1, 3, 12], телекомунікації [2, 3, 14] та сектор онлайн-ігор [16]. Водночас набагато менше уваги приділяється сектору роздрібної торгівлі (retail), детально розглянутий лише в [15] та поверхнево у [3], де відносини з клієнтами часто є неконтрактними, що робить завдання прогнозування відтоку та життєвої цінності клієнта ще складнішим і може вимагати специфічних підходів.

Остання вагома проблема пов'язана із розбіжністю між обсягом даних, які використовуються в дослідницьких експериментах, та тими, якими реально володіє бізнес. У наукових роботах тестування моделей часто відбувається на дуже багатих та різноманітних категоріях даних, що включають демографічні показники, детальну інформацію про сесії, графи соціальних зв'язків та історію звернень до служби підтримки [16]. Проте на практиці багато компаній, зокрема, в роздрібній торгівлі, можуть мати лише дуже обмежений набір даних, який зводиться виключно до базової інформації про транзакції (сума, дата, кількість товарів та ідентифікатор клієнта), від якого і потрібно відштовхуватися при побудові прогностичних систем [15]. Це створює значний розрив між теоретичними досягненнями багатокритеріальних моделей AutoML та їхньою практичною цінністю для компаній із базовим рівнем збору даних.

Зважаючи на вищезазначені прогалини, виникає гостра необхідність у дослідженні, яке б практично оцінило здатність сучасних систем AutoML автоматизувати роботу ML-інженерів в умовах обмежених даних у роздрібній торгівлі. Тому метою даної статті є оцінка ефективності передових фреймворків автоматизованого машинного навчання (на прикладі AutoSklearn та TROT) у вирішенні задачі прогнозування відтоку клієнтів. Для досягнення цієї мети буде виконано наступні кроки: підготовку відкритого транзакційного датасету, генерацію предметно-орієнтованих маркетин-

гових ознак, тестування моделей AutoML на різних часових лімітах із використанням методу історичної симуляції (out-of-sample testing), а також порівняння результатів роботи автоматизованих конвеєрів із ручною попередньою обробкою.

2. Підготовка даних

У даному дослідженні використовуються загальнодоступний транзакційний набір даних «Online Retail» (зокрема версія з репозиторію машинного навчання UCI [17]), який охоплює історію покупок клієнтів британського інтернет-магазину унікальних подарунків за дворічний період з 2009 по 2011 рік. Початковий масив даних налічує 1 067 371 рядок та 8 колонок із такою базовою інформацією: номер рахунку, унікальний код і назва товару, кількість куплених одиниць, дата та час транзакції, ціна товару у фунтах стерлінгів, унікальний ідентифікатор (ID) клієнта та країна транзакції.

Для підвищення точності моделювання масив даних пройшов комплексну попередню обробку. Процес очищення розпочався з видалення дублікатів та транзакцій без унікального ідентифікатора клієнта, що уможливило коректну агрегацію індивідуальної історії покупок. Згодом було вилучено скасовані замовлення (із префіксом «C») та нетипові технічні товари (наприклад, «ADJUST» чи «BANK CHARGES»), аби запобігти викривленню розрахунків доходу й поведінкових метрик. Вибірку географічно звузили до транзакцій із Великої Британії (понад 90% датасету), оскільки використання вітчизняних даних наразі неактуальне через суттєву зміну патернів споживання внаслідок війни. На фінальному етапі для стабілізації роботи алгоритмів було відсіяно статистичні викиди вище 99-го перцентилля, а всі валідні транзакції згруповано по днях, що дозволило сформувати чітку структуру щоденної активності кожного покупця.

Навіть за умови використання засобів AutoML, суттєво важливою залишається обробка та агрегація даних перед їх подачею в модель. Автоматизація не замінює необхідності у технічних навичках для

очищення масиву від дублікатів та аномалій, а також у знаннях предметної області для відсіювання нерелевантних записів чи врахування змін у поведінці. Якісний результат моделювання безпосередньо залежить від експертної підготовки вхідної вибірки, яка є фундаментом для будь-якого алгоритму.

3. Визначення відтоку

З технічної точки зору прогнозування відтоку клієнтів найчастіше формулюється як класична задача бінарної класифікації, де алгоритм машинного навчання має передбачити один із двох можливих станів: клієнт продовжить взаємодію з компанією або ж припинить її [3, 15]. Для коректної побудови та тренування такої прогностичної моделі життєвий цикл клієнта штучно розділяють на специфічні часові проміжки. Спочатку визначається вікно спостереження (*observation window*), протягом якого система агрегує історичні дані про транзакції, поведінку та будь-які взаємодії споживача з брендом для формування вхідних ознак [3, 12, 15]. За цим періодом слідує вікно прогнозування, яке також називають вікном маркування (*labeling window* або *evaluation window*). Саме в цьому вікні фіксується фактичний цільовий статус клієнта, а саме чи відбувся відтік. Часове та логічне розділення цих вікон є важливим для уникнення витоку даних, коли інформація з майбутнього могла б випадково потрапити у тренувальний набір і спотворити результати моделювання. Тривалість вікна маркування зазвичай залежить від специфіки конкретного бізнесу.

Саме поняття «відтоку клієнтів» є доволі розмитим і суб'єктивним, адже його визначення здатне суттєво відрізнитися залежно від типу послуг, галузі або стратегічних потреб конкретної компанії [12, 16]. У науковій літературі та комерційній практиці не існує єдиного універсального стандарту того, що саме вважати втратою клієнта. Наприклад, у банкінгу відтоком вважають зниження транзакцій на 30% або статус «неактивності» [16], тоді як в інших сферах критерієм є відсутність дій протягом 90 днів чи медіанного інтервалу між покуп-

ками [15]. Отже, формулювання задачі прогнозування майже завжди базується на евристичних правилах, які встановлюються експертами галузі або керівництвом для вирішення дуже специфічних поточних проблем бізнесу.

Через різноманітність цих визначень та характеристик різних ринків, нормальні показники відтоку суттєво коливаються. Наприклад, щорічно телекомунікаційний сектор втрачає 20–40% абонентів [2], банківський — 15–25% [8], а за іншими даними ці рівні становлять 20,4% та 26,5% відповідно [3]. Окремої уваги заслуговує сфера електронної та роздрібною торгівлі, де відносини між магазином та покупцем переважно не мають жодного контрактного характеру. У такому середовищі споживачі можуть дуже легко змінювати платформи та бренди, тому рівень відтоку тут традиційно є найвищим. Зокрема, дослідження на базі даних онлайн-магазину роздрібною торгівлі фіксують частку відтоку на рівні 32,1% [3].

Спираючись на проаналізовану інформацію, для подальшого дослідження було вирішено сфокусуватися лише на активних клієнтах, які зробили не менше 4 покупок упродовж усього часового періоду. Розмір оглядового вікна встановлений на рівні 12 місяців у зв'язку з тим, що специфіка даного датасету не передбачає дуже частих покупок (середня кількість днів між покупками становить 112 днів, а медіана — 82 дні). Вікно прогнозування визначене на рівні 6 місяців. Це дало можливість отримати датасет, у якому 22,4% клієнтів зазнали відтоку, що відповідає очікуванням з огляду на дані попередніх досліджень.

4. Підготовка ознак

Побудова власної прогностичної моделі відтоку клієнтів спирається на транзакційні дані для генерації набору поведінкових метрик, концептуально подібних до тих, що були розроблені у [15], головною відмінністю між якими виступає розмір оглядового вікна (1 рік замість 2 місяців). Основу набору ознак складають базові змінні: загальна кількість транзакцій, сумарні витрати клієнта, середній розмір чека за обраний період спостереження (який може

гнучко налаштовуватися залежно від бізнес-циклу), а також кількість днів, що минула з моменту останньої покупки. Цей вибір є обґрунтованим, оскільки такі метрики фактично є адаптацією класичної моделі RFM (давність, частота, грошова цінність), яка, згідно з численними дослідженнями, слугує галузевим стандартом для моделювання та прогнозування поведінки споживачів у багатьох індустріях [16]. Окрім цього, також включено метрику загальної тривалості активності клієнта, що допоможе алгоритмам оцінювати його загальну лояльність. Важливість цього кроку підтверджується ефективністю розширених моделей LRFM (до яких додається тривалість відносин, Length), які активно застосовуються в інших наукових працях для суттєвого покращення розпізнавання стабільних користувачів [16].

Крім статичних показників, важливим етапом є генерація розширених динамічних ознак, які фіксують відсоткові зміни в купівельній активності. Наслідуючи логіку [15], створено змінні для вимірювання відносного зростання або падіння кількості покупок та сум витрат порівняно з попередніми періодами, адже саме спад активності є найяскравішим раннім індикатором майбутнього відтоку. Прогностична цінність таких трендових індикаторів надійно підкріплена й іншими сучасними роботами. Наприклад, спеціалізовані системи автоматизованого машинного навчання, такі як Marketing-AutoM3L, за замовчуванням генерують подібні оцінки поведінкової залученості, обчислюючи швидкість і відносні тренди активності клієнтів для завчасного виявлення ризиків [3]. Такий же принцип доводить свою дієвість і у дослідженнях фінансового сектору, де аналіз спадних чи зростаючих трендів (нахилів) транзакційної активності користувачів дозволяв алгоритмам значно ефективніше та точніше прогнозувати відтік, ніж використання лише статичної демографічної інформації [12].

5. Масштабування й вибір ознак

Саме на цьому етапі порівнюється ефективність засобів автоматизованого машинного навчання із ручним підходом. За

основу ручного підходу взято дослідження [15], де для боротьби зі зміщеним розподілом значень, який часто виникає через екстремальні покупки клієнтів, застосовується логарифмічна трансформація. Після цього відбувається масштабування ознак за допомогою стандартизації Z-score, що приводить всі показники до єдиного діапазону. Щоб усунути проблему високої кореляції між згенерованими метриками, використовується метод ручного групування: виявляються ознаки із високою кореляцією, після чого їхні значення усереднюються, об'єднуються в єдиний показник. Це дозволяє зменшити розмірність даних і значно спростити роботу прогностичних алгоритмів.

На противагу ручному процесу, попередні системи автоматизованого машинного навчання, такі як AutoSklearn, пропонують широкий спектр вбудованих математичних методів для попередньої обробки. AutoSklearn повністю автоматизує масштабування, самостійно тестуючи та обираючи найоптимальніший метод серед стандартного, мінімаксного або робастного масштабування для конкретного набору даних. Щодо обробки сильно корельованих ознак, AutoSklearn пропонує прямий алгоритмічний аналог - метод агломерації ознак [18]. Цей алгоритм автоматично виконує кластеризацію змінних і об'єднує схожі ознаки, що по суті автоматизує процес ручного усереднення корельованих метрик. Крім того, AutoSklearn використовує потужні методи зменшення розмірності, такі як метод головних компонент (PCA) та швидкий аналіз незалежних компонент (Fast ICA), що дозволяє ефективно усувати надлишковість та шум без жодного ручного втручання експерта [18].

Інша потужна система, TPOT, підходить до масштабування та відбору ознак через призму генетичного програмування, будуючи гнучкі еволюційні конвеєри перетворень. Для масштабування TPOT автоматично інтегрує стандартні та робастні нормалізатори (StandardScaler, RobustScaler) безпосередньо у свої пайплайни, якщо еволюційний алгоритм підтверджує, що це підвищує загальну точність моделі [19]. Однак, на відміну від прямого групування та усереднення корельованих ознак, TPOT за-

стосовує рандомізований метод головних компонент (RandomizedPCA), рекурсивне виключення ознак (RFE) та фільтрацію за порогом дисперсії (Variance Threshold) [19]. Це означає, що замість злиття схожих метрик в одну, ТРОТ комбінує їх у нові нелінійні компоненти (головні компоненти) або відкидає ті з них, які еволюційний алгоритм визнає найменш корисними чи надто корельованими для кінцевого прогнозу.

6. Перебіг експерименту та його результати

У ході дослідження відбувалося тренування багатьох моделей машинного навчання, для яких було використано різні методи підготовки ознак та різні засоби AutoML для вибору й оптимізації моделей. Оскільки процес автоматизованого машинного навчання ресурсомісткий, для AutoSklearn і ТРОТ було надано різні проміжки часу з метою перевірити як часові рамки впливають на навчання моделей AutoML.

Експеримент полягає у проведенні порівняльного тестування AutoSklearn та ТРОТ у задачі прогнозування відтоку клієнтів. Оскільки процес автоматизованого машинного навчання ресурсомісткий, тренування було розподілене за різними часовими лімітами (5, 15 та 30 хвилин) і також за трьома підходами до підготовки даних: повністю ручне масштабування та групування ознак, лише ручне масштабування та використання повністю «сирих» маркетингових ознак.

Під час розробки та тренування моделей машинного навчання замість стандартного випадкового перемішування даних було застосовано підхід історичної симуляції, відомий як *out-of-sample historical testing* [15]. Цей метод передбачає навчання алгоритму на вибірці історичних даних з минулого та подальше тестування його прогнозів на абсолютно нових даних із майбутнього періоду. Важливість такого підходу полягає в тому, що реальні ринки та поведінка споживачів постійно змінюються у часі. Прогностичні моделі поводяться зовсім інакше в умовах простого випадкового розділення даних порівняно з реальним

прогнозуванням майбутніх подій. Завдяки історичній симуляції поза вибіркою можна отримати максимально реалістичну оцінку того, як саме модель працювала б, якби її запустили в комерційну експлуатацію в реальному часі.

Для забезпечення додаткової надійності та стабільності алгоритмів на етапі навчання також застосовано метод п'ятиблочної перехресної перевірки, тобто *5-fold cross-validation*, подібно до [1]. Цей підхід розділяє тренувальний масив даних на п'ять рівних частин, де кожна з них по черзі виступає в ролі валідаційного набору, тоді як інші чотири використовуються для навчання моделі. Використання перехресної перевірки є важливим, оскільки ефективно допомагає пом'якшити проблему перенавчання (*overfitting*) [16]. Оцінюючи модель на різних незалежних підмножинах даних, алгоритм узагальнює знайдені закономірності, а не просто заучує специфіку тренувальної вибірки.

Для оцінки прогнозів обрано комплекс метрик, що враховують значний класовий дисбаланс (22.4% відтоку). Оскільки стандартна точність (асигурація) у таких умовах може бути оманливою [16], основна увага приділена ROC AUC. Вона об'єктивно оцінює здатність моделі розрізняти класи незалежно від порогу відсічення [15]. Додатково використано *weighted F1* [14] — гармонійне середнє між точністю та повнотою, зважене за кількістю представників кожного класу. Для оцінки бізнес-ефективності застосовано *top-decile lift*. Цей показник демонструє перевагу моделі над випадковим вгадуванням серед 10% найбільш ризикових клієнтів, що дозволяє оптимізувати витрати на їхнє утримання [15, 16].

Отримані результати, наведені в таблиці, демонструють різну поведінку двох фреймворків за умов обмеженого часу. AutoSklearn показує високу стабільність і точність прогнозів усіх часових проміжків (*Weighted F1* на рівні 0.75–0.78, ROC AUC 0.783–0.792) незалежно від того, як були підготовлені дані. Така ефективність навіть за встановлених часових обмежень може бути зумовлена його архітектурою: система використовує байєсівську оптимізацію на базі випадкових лісів (SMAC),

яка суттєво пришвидшується завдяки мета-навчанню. Фреймворк аналізує метаознаки поточного набору даних, використовує досвід попередніх успішних оптимізацій для «теплого старту», а наприкінці автоматично поєднує найкращі знайдені моделі в ефективний ансамбль.

На противагу цьому, продуктивність TPOT суттєво залежить від наданого йому обчислювального часу. За умови ліміту в 5 хвилин на «сирих» даних його результати є

часу ручне масштабування дещо допомогло TPOT знайти рішення швидше (Accuracy 0.710 на 5 хвилинах), але не розкрило його повного потенціалу.

Ще одним показовим висновком з експерименту є вплив ручної попередньої обробки порівняно з автоматизованою. Подача «сирих» (немасштабованих і незгрупованих) даних може несуттєво погіршити деякі результати (падіння ROC AUC з 0.799 до 0.792), а за достатнього часу тренування

AutoML tool	AutoSklearn				TPOT			
Testing on features scaled and grouped manually	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift
5m	0.785	0.783	0.75	2.390	0.710	0.736	0.73	2.148
15m	0.791	0.794	0.75	2.685	0.786	0.787	0.74	2.497
30m	0.786	0.794	0.74	2.470	0.785	0.799	0.74	2.551
Testing on features scaled manually	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift
5m	0.789	0.790	0.75	2.578	0.711	0.785	0.73	2.470
15m	0.784	0.796	0.72	2.578	0.726	0.789	0.75	2.524
30m	0.779	0.784	0.69	2.685	0.781	0.792	0.71	2.470
Testing on raw features	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift	Accuracy	ROC AUC	Weighted F1	Top-Decile Lift
5m	0.798	0.792	0.78	2.658	0.647	0.745	0.66	2.739
15m	0.787	0.778	0.78	2.658	0.793	0.791	0.78	2.524
30m	0.787	0.769	0.78	2.685	0.800	0.787	0.78	2.497

менш показовими (Accuracy лише 0.647, Weighted F1 падає до 0.66), проте зі збільшенням часу до 15 та 30 хвилин він різко стабілізує роботу, досягаючи рівня AutoSklearn, іноді й перевершуючи показники з невеликим відривом (Accuracy зростає до 0.800, а Weighted F1 до 0.78). Ця динаміка відповідає закладеним у TPOT алгоритмам генетичного програмування, де система ітеративно будує, мутує та схрещує деревоподібні конвеєри перетворень. Мутація пайплайнів із подальшою багаточисловою Парето-оптимізацією є вкрай ресурсомістким процесом, що вимагає значного часу для еволюції моделі та збіжності до оптимального рішення. За обмеженого

дозволила обом фреймворкам досягти найвищих показників Weighted F1 (0.78) у порівнянні з ручною обробкою (0.71–0.75). Це підтверджує тезу про те, що сучасні засоби AutoML здатні на рівні з людиною автоматизувати процеси нормалізації та боротьби з мультиколінеарністю. AutoSklearn самостійно тестує різні методи масштабування та використовує алгоритм агломерації ознак або аналіз головних компонент (PCA) для усунення надлишковості та шуму. Своєю чергою, TPOT інтегрує нормалізатори безпосередньо у свої еволюційні пайплайни, а замість ручного злиття схожих метрик комбінує їх у нові нелінійні компоненти або алгоритмічно відкидає через ре-

курсивне виключення ознак (RFE). Таким чином, ручне логарифмування, стандартизація Z-score та логічне усереднення залежних змінних виявляються зайвими та певною мірою обмежувальними кроками.

У контексті бізнес-аналітики, оцінювати ці моделі варто через призму профільних метрик. Через значний дисбаланс класів базова точність (Ассигасу) може бути зміщеною та оманливою. Значно важливішими є стабільні показники ROC AUC та weighted F1, які успішно тримаються на рівні близько 0.78–0.79. Водночас метрика Top-Decile Lift, яка досягає значень 2.390–2.739, оозначає, що навчені моделі AutoML дозволяють виявляти клієнтів із найвищим ризиком відтоку (у верхніх 10% клієнтської бази) у приблизно два з половиною рази ефективніше, ніж якби маркетологи обирали аудиторію навмання, що демонструє суттєву комерційну цінність побудованих конвеєрів. Завдяки цьому бізнес може максимально раціонально оптимізувати маркетинговий бюджет, спрямовуючи пропозиції з утримання виключно на тих споживачів, які цього справді потребують.

Результати підтверджують, що для швидкого отримання надійних прогнозів на непідготовлених даних добре підходить AutoSklearn, тоді як TPOT потребує значних обчислювальних витрат і часу. Однак важливо розуміти, що високі результати обох систем стали можливими також завдяки тому, що до їхнього запуску було проведено якісну інженерію доменних ознак: вхідні дані вже містили розраховані маркетингові RFM-метрики, тривалість активності рахунків та індикатори поведінкової залученості клієнтів. Засоби AutoML беруть на себе математичну рутину обробки та відбору ознак, але формування правильного маркетингового контексту та створення релевантних змінних все ще лежить в основі успішного вирішення проблеми.

Висновки

У статті запропоновано та практично оцінено ефективність впровадження засобів автоматизованого машинного навчання (AutoML) для мінімізації ручного втручання технічних спеціалістів у процес по-

будови предиктивних моделей на базі обмежених наборів даних. Проведені експерименти підтвердили основну гіпотезу дослідження, згідно з якою сучасні платформи (AutoSklearn та TPOT) здатні успішно та на рівні з людиною автоматизувати процеси нормалізації, масштабування та боротьби з мультиколінеарністю без втрати якості кінцевого прогнозу. У межах роботи формалізовано задачу прогнозування відтоку клієнтів у неконтрактному середовищі роздрібної торгівлі, де бізнес оперує виключно базовою транзакційною історією. Валідація алгоритмів здійснювалася методами історичної симуляції (out-of-sample testing) та 5-блочної перехресної перевірки з використанням 11 розроблених поведінкових метрик.

Експериментальна оцінка продемонструвала високу ефективність запропонованого підходу: подача «сирих» (немасштабованих і незгрупованих) даних у системи AutoML дозволила досягти найвищих показників Weighted F1 (0.78) порівняно з їхньою попередньою ручною обробкою (0.71–0.75). Аналіз часових лімітів виявив, що AutoSklearn здатен генерувати стабільні конвеєри з ROC AUC на рівні 0.792 вже за 5 хвилин обчислень завдяки механізмам метанавчання, тоді як архітектура TPOT потребувала від 15 до 30 хвилин для збіжності до аналогічних результатів. Побудовані моделі забезпечили показник Top-Decile Lift до 2.739, що дозволяє виявляти ризикових клієнтів у 2,5 рази ефективніше за випадковий вибір. Отримані результати підтверджують доцільність використання фреймворків AutoML для розгортання прогностичних ІТ-систем в умовах дефіциту розширених даних та нестачі часу технічних експертів, водночас визначаючи ключову роль якісної генерації доменних ознак перед етапом автоматизації.

References

1. S. Dao, T. Dong, S. Chen, Automated Customer Churn Prediction in Banking: A Domain-Aware AutoML Approach (2026).
2. M. Mandić, G. Kraljević, Churn prediction model improvement using automated machine learning with social network parameters,

- Revue d'Intelligence Artificielle 36 (3) (2022) 373–379. doi: 10.18280/ria.360304
3. Y. Tian, W. Shao, Z. Deng, Marketing-AutoM3L: domain-aware automated machine learning for financial customer analytics, *Frontiers in Artificial Intelligence* 9 (2026) 1726900. doi: 10.3389/frai.2026.1726900
 4. M.-A. Zöllner, M.F. Huber, Benchmark and Survey of Automated Machine Learning Frameworks, *Journal of Artificial Intelligence Research* 70 (2019) 409–472.
 5. D. Luo, C. Feng, Y. Nong, Y. Shen, AutoM3L: An Automated Multimodal Machine Learning Framework with Large Language Models, in: *Proceedings of the 32nd ACM International Conference on Multimedia* (2024).
 6. P. Trirat, W. Jeong, S.J. Hwang, AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML, *arXiv preprint arXiv:2410.02958* (2024).
 7. X. He, K. Zhao, X. Chu, AutoML: A Survey of the State-of-the-Art, *arXiv preprint arXiv:1908.00709* (2019).
 8. A. Mari, *The Rise of Machine Learning in Marketing: Goal, Process, and Benefit of AI-Driven Marketing* (2019).
 9. D. Herhausen, S.F. Bernritter, E.W.T. Ngai, A. Kumar, D. Delen, Machine learning in marketing: Recent progress and future research directions, *Journal of Business Research* 170 (2024) 114254. doi: 10.1016/j.jbusres.2023.114254
 10. M.S. Kasem, M. Hamada, I. Taj-Eddin, Customer profiling, segmentation, and sales prediction using AI in direct marketing, *Neural Computing and Applications* 36 (2024) 4995–5005. doi: 10.1007/s00521-023-09339-6
 11. B. Gao, Y. Wang, H. Xie, Y. Hu, Y. Hu, Artificial Intelligence in Advertising: Advancements, Challenges, and Ethical Considerations in Targeting, Personalization, Content Creation, and Ad Optimization, *Sage Open* 13 (4) (2023).
 12. E. Kaya, X. Dong, Y. Suhara, et al., Behavioral attributes and financial churn prediction, *EPJ Data Science* 7 (2018) 41. doi: 10.1140/epjds/s13688-018-0165-5
 13. M. Mandić, G. Kraljević, Two-Layer Architecture of Telco Churn Auto-ML (2020).
 14. M.V.C. Aragão, A.G. Afonso, R.C. Ferraz, et al., A practical evaluation of AutoML tools for binary, multiclass, and multilabel classification, *Scientific Reports* 15 (2025) 17682. doi: 10.1038/s41598-025-02149-x
 15. S. Akhmetbek, Forecasting Customer Future Behavior in Retail Business Using Machine Learning Models, *Scientific Journal of Astana IT University* (2022).
 16. A. Manzoor, M.A. Qureshi, E. Kidney, L. Luca, A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners, *IEEE Access* 12 (2024) 70434–70463.
 17. D. Chen, Online Retail II (Version 1) [Data set], *UCI Machine Learning Repository* (2019).
 18. M. Feuerer, K. Eggensperger, S. Falkner, M. Lindauer, F. Hutter, Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning, *arXiv preprint arXiv:2007.04074* (2020).
 19. P. Ribeiro, et al., TPOT2: A New Graph-Based Implementation of the Tree-Based Pipeline Optimization Tool for Automated Machine Learning, in: S. Winkler et al. (Eds.), *Genetic Programming Theory and Practice XX*, Springer, Singapore (2024). doi: 10.1007/978-981-99-8413-8_1

Дата першого надходження до видання:
07.03.2026

Внутрішня рецензія отримана: 14.03.2026

Зовнішня рецензія отримана: 14.03.2026

Дата прийняття статті до друку: 19.03.2026

Дата публікації: 16.04.2026

Про автора:

¹Ніконов Олександр Володимирович,
аспірант

Nikonov Olexandr,

Post-graduate student

<https://orcid.org/0009-0009-4743-4854>.

Місце роботи автора:

¹ Національний технічний університет
України «Київський політехнічний
інститут імені Ігоря Сікорського»
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
E-mail: nikonov.sanynikonov@gmail.com
Сайт: <https://ist.kpi.ua>