ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ДЛЯ ПРОГНОЗУВАННЯ У БАНКІВСЬКІЙ СФЕРІ

Міщенко А.С.[1], Гуськова В.Г.

Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна

[1] mischenkoanton7@gmail.com

У роботі розглянуто підхід до формування та прогнозування клієнтського портфелю банку на основі методів машинного навчання та інтелектуальних систем підтримки прийняття рішень. Наведено етапи підготовки даних, побудови ознак, сегментації клієнтів за допомогою алгоритму KMeans та розробки моделей класифікації для прогнозування ймовірності відкриття депозитів і отримання кредитів. Отримані результати демонструють можливість підвищення точності прогнозів порівняно з традиційними підходами.

Ключові слова: банківська аналітика, прогнозування, машинне навчання, сегментація, клієнтський портфель, інтелектуальні системи.

1. ВСТУП

У сучасних умовах цифрової трансформації банківського сектору питання формування та управління клієнтським портфелем набуває особливої актуальності. Банки працюють з великими масивами даних про клієнтів, їх фінансову поведінку, історію взаємодії з продуктами та каналами обслуговування. Ефективний аналіз цих даних дозволяє не лише підвищувати прибутковість, але й зменшувати ризики, пов’язані з неповерненням кредитів, відтоком клієнтів або неефективним використанням ресурсів.

Традиційні статистичні методи часто виявляються недостатньо гнучкими для роботи з високовимірними та гетерогенними даними, що характерні для клієнтських баз банків. У зв’язку з цим зростає роль інтелектуальних систем та алгоритмів машинного навчання, які здатні автоматично виявляти приховані закономірності, сегментувати клієнтів за поведінковими ознаками та прогнозувати їх подальші дії. Особливого значення набувають моделі класифікації, що дозволяють оцінювати ймовірність відкриття депозиту чи залучення кредиту в розрізі кожного клієнта.

У рамках даної роботи розглядається програмна реалізація повного циклу побудови системи прогнозування клієнтського портфелю – від підготовки вихідних даних до сегментації клієнтів, навчання моделей та інтерпретації отриманих результатів.

2. ПОСТАНОВКА ЗАДАЧІ ТА ВИХІДНІ ДАНІ

2.1. Мета та задачі дослідження

Метою дослідження є розробка та використання інтелектуального інструментарію для аналізу та прогнозування клієнтського портфелю банку на основі методів машинного навчання. Для досягнення поставленої мети необхідно вирішити такі основні задачі: визначити релевантні вхідні ознаки, забезпечити коректну підготовку даних, побудувати сегментаційну модель клієнтської бази та розробити класифікаційні моделі для прогнозування цільових показників.

У роботі розглядаються дві ключові задачі прогнозування. Перша пов’язана з оцінюванням ймовірності відкриття клієнтом строкового депозиту, друга – з прогнозуванням ймовірності оформлення кредиту. Обидві задачі мають вигляд бінарної класифікації, де результатом є належність клієнта до однієї з двох груп: «буде користуватися продуктом» або «не буде». Такі прогнози є важливими для планування ресурсів банку, таргетованих маркетингових кампаній та управління ризиками.

2.2. Характеристика вихідних даних

Вихідні дані являють собою табличну вибірку, що містить інформацію про соціально-демографічні характеристики клієнтів, параметри банківських продуктів та історію транзакційної активності. До набору ознак входять, зокрема, вік, сімейний стан, рівень доходів, наявність поточних кредитів, тип платіжної картки, рівень використання кредитного ліміту, частота та обсяги операцій. Окремі стовпці відповідають цільовим змінним – факту наявності строкового депозиту та факту користування кредитними продуктами.

Оскільки в реальних банківських системах особлива увага приділяється захисту персональних даних, при підготовці вибірки було закладено принципи анонімізації клієнтських ідентифікаторів. Для цього створено спеціальне поле з маскованим ідентифікатором клієнта, в якому зберігаються лише перші та останні символи, а проміжні замінюються на «*». Такий підхід дозволяє з одного боку зберегти можливість відстежувати записи одного клієнта, а з іншого – унеможливлює ідентифікацію конкретної особи.

3. МЕТОДИ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1. Попередня обробка та побудова ознак

Попередня обробка даних включає заповнення пропусків, кодування категоріальних змінних та масштабування числових ознак. Для числових полів застосовано медіанне заповнення пропусків та стандартизацію, що забезпечує однаковий масштаб змінних і сприяє стабільнішій роботі моделей. Категоріальні ознаки було перетворено за допомогою one-hot кодування, при цьому використано обмеження на максимальну кількість категорій, що дозволяє уникнути надмірного розширення простору ознак.

Особливу увагу приділено виключенню змінних, які можуть призводити до витоку інформації про цільову змінну. Зокрема, з набору ознак було вилучено технічні поля, що напряму або опосередковано містять підказку щодо факту відкриття продукту або результатів минулих маркетингових кампаній. Такий підхід дає змогу отримати більш реалістичну оцінку якості моделей та підвищити їх придатність до використання у виробничих умовах.

3.2. Сегментація клієнтів методом KMeans

Для сегментації клієнтів використано алгоритм KMeans, який дозволяє розбити вибірку на однорідні кластери за сукупністю поведінкових та соціально-демографічних ознак (рис. 1). Перед сегментацією дані було перетворено за допомогою побудованого препроцесора, що забезпечило коректний масштаб ознак та усунення впливу різниці в одиницях виміру. Оптимальну кількість кластерів визначено на основі аналізу коефіцієнта силуету та графіка «лікоть», які відображають баланс між компактністю кластерів та відстанню між ними.

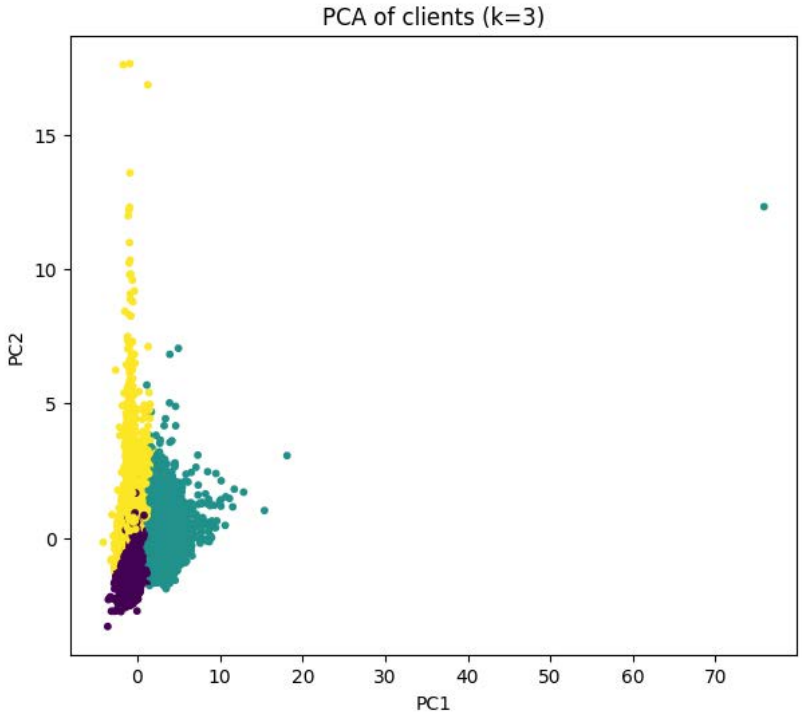

Рисунок 1. Візуальне відображення сегментації клієнтів

У результаті було отримано декілька сегментів, які відрізняються за середнім рівнем доходів, віковою структурою, активністю користування продуктами та схильністю до депозитних чи кредитних операцій. Наприклад, один із сегментів характеризується молодшою аудиторією з невисокими доходами, але високою транзакційною активністю, тоді як інший сегмент включає клієнтів середнього віку з вищими доходами та інтенсивним використанням кредитних продуктів (табл. 1). Таке профілювання дозволяє банку більш точно налаштовувати маркетингові стратегії.

Таблиця 1. Сегментація клієнтів

cluster	age	balance	day	campaign
0	34.0606	941.32	16.09	2.82
1	39.4215	1324.66	13.93	2.13
2	51.4592	1923.72	16.22	3.06

3.3. Побудова моделей прогнозування

Для розв'язання задач прогнозування було обрано декілька моделей машинного навчання, серед яких логістична регресія, випадковий ліс (Random Forest) та градієнтний бустинг. Навчання моделей здійснювалося за допомогою конвеєрів (Pipeline), що поєднують у собі препроцесор ознак та безпосередньо алгоритм класифікації. Такий підхід спрощує повторне навчання на нових даних і зменшує ризик помилок при ручному виконанні окремих етапів.

Якість моделей оцінювалася за стандартними метриками бінарної класифікації: точністю, повнотою, F1-мірою та площею під ROC-кривою (табл. 2). Результати експериментів показали, що ансамблеві методи, зокрема випадковий ліс та градієнтний бустинг, забезпечують кращий компроміс між чутливістю та специфічністю порівняно з базовою логістичною регресією. Отримані значення F1-міри та ROC-AUC свідчать про можливість практичного використання моделей для підтримки прийняття рішень у банку.

Таблиця 2. Порівняння показників точності моделей

	task	model	accuracy	F1	precision	recall	Roc_auc	y_test_mean
0	B_term_deposit	LogReg	0.8993	0.3469	0.4364	0.2878	0.7217	0.0928
1	B_term_deposit	RandomForest	0.9104	0.1450	0.6377	0.0818	0.7372	0.0928
2	B_term_deposit	GradBoost	0.9112	0.1858	0.6279	0.1090	0.7472	0.0928
3	C_take_credit	LogReg	0.7720	0.8218	0.8251	0.8186	0.8270	0.6424
4	C_take_credit	RandomForest	0.7852	0.8398	0.8060	0.8767	0.8400	0.6424
5	C_take_credit	GradBoost	0.7883	0.8433	0.8041	0.8865	0.8492	0.6424

4. РЕЗУЛЬТАТИ СЕГМЕНТАЦІЇ ТА ПРОГНОЗУВАННЯ

4.1. Характеристика отриманих сегментів

На основі результатів роботи алгоритму KMeans було виділено кілька стійких сегментів клієнтів, для яких розраховано середні значення ключових числових показників та визначено домінуючі категорії за якісними ознаками. Це дозволило сформувати узагальнені «портрети» клієнтів для кожного сегменту. Зокрема, один сегмент об’єднує молодь з відносно невисокими доходами, яка активно використовує платіжні картки для повсякденних операцій, але рідше користується кредитами. Інший сегмент представлений клієнтами середнього віку з вищими доходами, високою кредитною активністю та схильністю до відкриття депозитів.

Ще один сегмент формується за рахунок більш старших клієнтів, які мають стабільні доходи та значні залишки на рахунках, але демонструють низьку транзакційну активність. Для цієї групи більш релевантними є продукти, пов’язані зі збереженням та нарощенням капіталу. Таким чином, сегментація дозволяє ідентифікувати групи з різною чутливістю до депозитних та кредитних пропозицій, що є важливим для таргетованого маркетингу.

4.2. Інтерпретація прогнозних результатів

Після навчання моделей для кожного клієнта було розраховано ймовірність відкриття депозиту та отримання кредиту. На основі показників точності кожної з моделей, які були використані, була обрана з найкращими результатами. Додатково було проаналізовано розподіл прогнозів у розрізі сегментів, що дозволило визначити, які кластери є найбільш перспективними для просування певних продуктів.

4.3. Результати роботи

Отриманим результатом роботи є сформована сегментована база клієнтів із зазначенням імовірності того, що кожен клієнт відкриє депозит чи оформить кредит (табл. 3). Для кожного сегменту система автоматично генерує перелік клієнтів із найвищими прогнозними значеннями, що дозволяє банку оперативно визначати найбільш перспективні цільові групи. Такий підхід спрощує аналіз великих вибірок і забезпечує можливість подальшої персоналізації пропозицій для кожної групи. На основі отриманих прогнозів банк може ефективно планувати маркетингові кампанії та підвищувати результативність комунікації з клієнтами.

Таблиця 3. Приклад результату роботи програми

	Client_id_masked	Proba_deposit
34017	217****5931	0.9475
31283	385****0760	0.9350
33953	798****0249	0.9275
34047	653****6346	0.9250
39763	363****5578	0.9225

5. ВИСНОВКИ

У роботі представлено підхід до побудови інтелектуальної системи аналізу та прогнозування клієнтського портфелю банку, що поєднує методи попередньої обробки даних, сегментації та машинного навчання. Запропоновано структуру програмної реалізації, яка включає модулі побудови ознак, сегментації клієнтів методом KMeans та навчання моделей класифікації для прогнозування ймовірності відкриття депозитів і отримання кредитів.

Результати експериментальних досліджень показали, що використання ансамблевих алгоритмів, таких як випадковий ліс та градієнтний бустинг, забезпечує високу якість класифікації та може бути ефективно застосоване у банківській практиці. Сегментація клієнтів дозволяє виділити групи з різним рівнем схильності до використання певних продуктів, що створює додаткові можливості для таргетованого маркетингу та оптимізації ризиків.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Breiman L. Random Forests. Machine Learning. 2001. Vol. 45, No. 1. P. 5–32.
2. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 2001. Vol. 29, No. 5. P. 1189–1232.
3. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
4. Hosmer D.W., Lemeshow S., Sturdivant R.X. Applied Logistic Regression. 3rd ed. Wiley, 2013.
5. Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 2005.
6. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65.
7. Hair J.F. et al. Multivariate Data Analysis. 8th ed. Cengage Learning, 2019.
8. Box G.E.P. et al. Time Series Analysis: Forecasting and Control. 5th ed. Wiley, 2016.
9. Gujarati D.N., Porter D.C. Basic Econometrics. 5th ed. McGraw-Hill, 2009.
10. Hand D.J., Mannila H., Smyth P. Principles of Data Mining. MIT Press, 2001.