УДК 339.138:338.27
DOI: https://doi.org/10.32782/2415-8801/2020-5.19

Зомчак Л.М.
кандидат економічних наук,
доцент кафедри економічної кібернетики,
Львівський національний університет імені Івана Франка

Вдовин М.Л.
кандидат економічних наук,
доцент кафедри статистики,
Львівський національний університет імені Івана Франка

ПРОГНОЗУВАННЯ УСПІШНОСТІ БАНКІВСЬКОГО МАРКЕТИНГУ МЕТОДАМИ ЛОГІСТИЧНОЇ РЕГРЕСІЇ

У статті дано модель логістичної регресії, побудованої на основі статистичної бази про результати опитування користувачів банківських послуг засобами телефонного зв'язку та електронної пошти. Логістична регресія як метод класичного машинного навчання із учителем дозволив класифікувати споживачів банківських послуг на тих, для котрих маркетингові заходи у формі електронних листів та дзвінків ефективні, та на тих, для котрих такі маркетингові заходи недієві. На основі розробленої моделі можна прогнозувати ефективність банківського маркетингу для конкретного споживача банківських послуг, при чому результуюча змінна показує ймовірність належності споживача до однієї із груп. Модель може бути використана банками для прогнозування успішності маркетингових кампаній, оптимізації цілей в межах таргетингових груп користувачів банківських послуг під час розроблення маркетинговий кампаній, а також ухвалення ефективних управлінських рішень у банківській сфері на основі даних.

Ключові слова: *логістична регресія, маркетинг, прогноз, банк, споживач.*

FORECASTING BANKING MARKETING SUCCESS WITH LOGISTIC REGRESSION METHODS

Zomchak Larysa, Vdovyn Mariana
Ivan Franko National University of Lviv

The aim of the article is to investigate the impact of telephone calls (so called telemarketing) and e-mail surveys on the purchase of banking services by consumers. As the part of the marketing campaign, the bank's employees call their customers and, at the same time, potential consumers of new services, and offer to open a deposit; The client receives a similar offer if he contacts the contact center of the bank for any other reason. Accordingly, the result of the marketing campaign is binary: successful or unsuccessful contact, i.e. the client put money on deposit or did not. Based on the available statistical information on conversations with the bank's clients, it is necessary to build a classification model that determines whether the client will make a deposit based on his characteristics. For binomial logistic regression construction, statistical data, that characterize potential consumers of the product were used, namely their age, gender, level of education, marital status, position at work, etc. The model is implemented in the R-Studio environment. The logistic regression model is realized on the basis of a statistical database, which describing users of banking services, who responded by telephone and e-mail. The basic idea of logistic regression is to use an already developed linear regression mechanism by adding probability, using a linear prediction function that explains variables and a set of regression coefficients that are unique to the model but the same for

all iterations. Logistic regression as a method of classical machine learning with a teacher allowed to classify consumers of banking services for those sensitive to marketing calls and for those for whom such marketing activities are ineffective. Based on the developed model, it is possible to predict the effectiveness of banking marketing for specific monitoring of banking services, so that the effective change shows that the level of reliability of consumers to the same group. The model can be used by the bank to predict the success of marketing projects, optimize prices within the target groups of banking users during the development of marketing services, as well as increase effective management decisions in the banking sector on dream data.

Keywords: *logistic regression, marketing, forecast, bank, consumer.*

Постановка проблеми. Швидкий та легкий доступ споживачів до інформації про товари, пов'язаний із поширенням її в електронному вигляді, зумовив більш обґрунтований вибір споживачів та більш ретельний підхід споживачів до покупок. Найчастіше головними критеріями споживача є поєднання високої якості товару та низької ціни. Однак, зазвичай, споживачі не можуть чітко обґрунтувати, чому віддають перевагу одним товарам перед іншими, чим керуються під час вибору, чи піддаються впливу рекламних пропозицій. Банківський маркетинг також зазнав значних змін під впливом цих тенденцій. Розроблення успішних банківських маркетингових стратегій передбачає застосування підходів, які базуються на даних, котрі, в свою чергу, реалізовуються методами машинного навчання.

Аналіз останніх досліджень і публікацій. Впродовж останніх років з'явилось багато публікацій, присвячених застосуванню методів машинного навчання та кількісного аналізу даних в маркетингу. Так, методам побудови рекомендаційних маркетингових систем на основі даних присвячені публікації Негрей М. та Гнота Т. [1; 2]. Х. Есаламони [3], Л. Сінговей [4] та Р. Паллар [5] досліджують можливості застосування методів інтелектуального аналізу даних у прямому банківському маркетингу.

Загальний огляд можливостей застосування методів машинного навчання у банківському маркетингу можна знайти у К. Ванга [6]; С. Гош та співавтори провели порівняльний аналіз методів прогнозування на основі даних банківського телемаркетингу [7].

Методи класифікації на основі машинного навчання, їх порівняння, переваги та недоліки розглянуті у статті К. Вісаенга [8]. В. Гронка [9] зі співавторами розглядають можливості застосування методів класифікації з учителем, які базуються на деревах рішень та їхніх модифікаціях, а С. Янг та Т. Чен методи невизначених дерев рішень [10]. Т. Янг зі співавторами реалізовують модель наївного баєсівського класифікатора та асоціативних правил на основі даних, отриманих за результатами маркетингової кампанії в банку [11]. К. Крішна та П. Редді застосовують для розв'язування задачі класифікації даних банківського маркетингу метод глибоких нейронних мереж та порівнюють із результатами, отриманими за допомогою чотирьох стандартних класифікаторів [12]. Шість методів машинного навчання, а саме: методи дерева рішень, наївний баєсівський класифікатор, нейронні мережі, метод опорних векторів, логістична регресія та випадкові ліси реалізовано у статті А. Верма [13]. Із аналізу літератури зрозуміло, що класифікація методами машинного навчання широко застосовується до маркетингових даних банків, при чому вибір методу залежить значною мірою як від цілі дослідження, так і від особливостей вибірки.

Постановка завдання. Необхідно дослідити вплив телефонних дзвінків та опитувань на здійснення покупки користувачами банківських послуг. У межах маркетингової кампанії працівники банку телефонують своїм клієнтам та, водночас, потенційним споживачам нових послуг, і пропонують відкрити депозит; аналогічну пропозицію клієнт отримує, якщо сам зв'язується з контакт-центром банку з будь-якої іншої причини. Відповідно результат бінарний: успішний або неуспішний контакт, тобто закладений чи ні депозит клієнтом. На основі наявної статистичної інформації про розмови із клієнтами банку потрібно побудувати класифікаційну модель, яка дозволить визначати, чи закладе клієнт депозит на основі його характеристик.

Виклад основного матеріалу дослідження. Дискретні моделі вибору дозволяють спрогнозувати вибір споживача між двома чи більше товарами, а найпоширенішою серед цього класу моделей є логістична модель.

Логістична регресія – це тип кривої регресії $y = f(x)$, де y – категорійна змінна. У типовому випадку для результуючої змінної y заданий набір факторних змінних x, при чому факторні змінні можуть бути як категорійними, так і кількісними.

Припущення, які передують застосуванню логістичної регресії, аналогічні до тих, які висувають до лінійної регресії. Однак, на відміну від лінійної регресії, на виході логістичної регресії отримують не прогноз змінної, а значення функції ймовірності належності до певного класу.

Оскільки при прогнозуванні на основі методів машинного навчання головна мета полягає в отриманні точних прогнозів, а не в інтерпретації результатів, то допустиме порушення деяких припущень.

Нехай x – факторна змінна, y – результуюча змінна, тоді вони утворюють точку з координатами (x, y). Модель логістичної регресії матиме вигляд:

$$y = logit^{-1}(z) + \varepsilon = \frac{1}{1+\exp(-z)} + \varepsilon,$$

де $z = b_0 + \sum_{j=1}^{n} b_j x_j$.

Нехай $p_i = (b, x_i)$, вектор $b = [b_0 \ldots b_n]^T$. Позначимо вибірку факторних змінних як:

$$X = \begin{matrix} 1 & x_1^T \\ .. & .. \\ 1 & x_m^T \end{matrix}.$$

Потрібно знайти таке значення вектора b, щоб виконувалася умова:

$$S = \|y - p\|^2 = \sum_{i=1}^{m} (y_i - p_i)^2.$$

За реальних умов проведення маркетингової кампанії постійно додаються нові дані у вибірку, тому, для досягнення надійної прогнозної оцінки застосовуємо схему оцінки типу «обертового вікна» з фіксованим

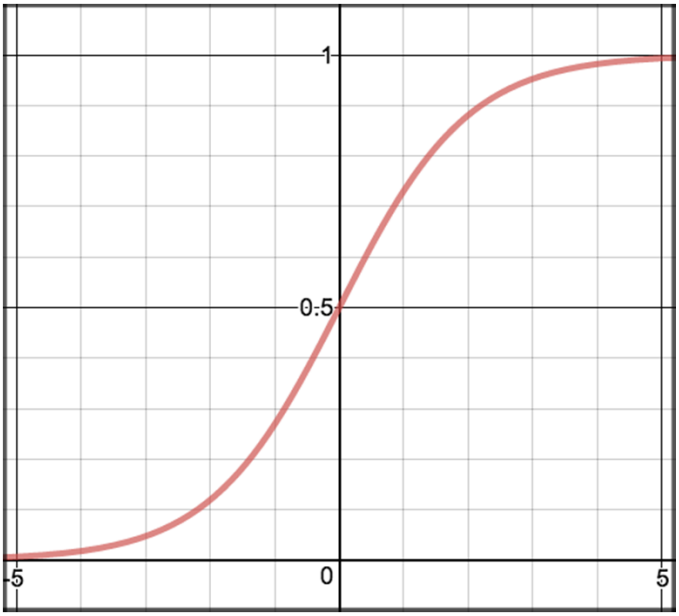

Рис. 1. Графік можливих значень вихідних параметрів функції

розміром, яка виконує оновлень моделей і відкидає найстаріші дані.

Традиційно, дані вхідної статистичної вибірки розподілимо за часом на дві частини:

1. Тренувальні дані (training data) – за період 4-х років;
2. Тестові дані (test data) – за останній рік.

Існують різні способи розбиття вибірки: проста випадкова вибірка і стратифікована вибірка на підставі результату, за датою та методами тощо. Скористаємось способом випадкової вибірки даних.

Тренувальні дані використовують для вибору функції і моделі. Кожен із записів містить інформацію про результат розмови (тобто успіх або провал) та вхідні характеристики споживача. До характеристик споживача належать: атрибути телемаркетингу (наприклад, напрямок дзвінка), деталі продукту (наприклад, процентна ставка) та інформація про клієнта. Також до цих записів додано дані загального соціального та економічного характеру, шляхом збору зовнішніх даних.

Для моделювання за допомогою R був використаний «caret» пакет.

На першому етапі імпортуємо та очищуємо дані. Зазвичай частина очищення даних займає багато часу і передбачає кілька етапів, таких як виправлення помилок форматування, виправлення відсутніх або помилкових значень та стандартизація категорійних стовпців. В цьому прикладі етап очищення даних був пропущений, оскільки всі дані були ретельно відібрані перед цим.

Другий етап – розподіл тренувальні та тестові дані.

Мета створення прогнозної моделі полягає у прогнозуванні нових даних. Якщо використовувати всі дані, доступні для створення моделі, не зможемо перевірити продуктивність моделі з набором даних, який не використовувався для його створення. Тому правильно розбити початкові набори даних у навчальні та тестові набори. Тренувальний набір буде використаний для створення моделі, а тестовий набір для оцінки продуктивності моделі. Розподіл задано наступним чином: 70% початкових даних використано для тренувального набору, інші 30% для тестового набору

Етап третій – створення моделі.

Для того, щоб перетворити прогнозні значення до ймовірностей використовуємо сигмоїдну функцію. Ця функція перетворює будь-яке дійсне значення на значення в діапазоні від 0 до 1 (рис. 1). У машинному навчанні сигмоїд використовується для перетворення прогнозів до ймовірностей.

$$S(z) = \frac{1}{1+e^{-z}},$$

де $s(z)$ = вихідний параметр в діапазоні від 0 до 1 (оцінка ймовірності);

z – вхідне значення (передбачення вашого алгоритму);

e – основа натурального логарифму.

На рисунку 1 представлений графік можливих значень вихідних параметрів функції.

Отримуємо прогноз для заданого набору даних, який показує чи погодиться клієнт на депозит.

Поточна функція прогнозування повертає оцінку ймовірності між 0 і 1. Для того, щоб віднести цю оцінку до дискретного класу, вибираємо граничне значення і припускаємо, якщо значення оцінки більше за обране граничне значення, будемо класифікувати його в класі 1 і якщо нижче – ми класифікуємо значення в клас 2:

$$p \geq 0.5, class = 1;$$
$$p \leq 0.5, class = 0.$$

Коли є набір прогнозів, для оцінювання ефективності моделі можна використовувати різні показники.

Для моделей регресії:

– середня квадратична похибка;
– коефіцієнт детермінації;
– кореляція Спірмена.

Для класифікаційних моделей:

– показник загальної точності, але коли класи є незбалансовані це може викликати певні труднощі;

– статистика Каппа враховує очікувану частоту помилок.

Тепер, маючи прогноз, отриманий за допомогою тренувального набору даних, можемо оцінити продуктивність моделі на наборі тестових даних. Найбільш базовим показником продуктивності є загальна точність, тобто частка випадків, коли результати спрогнозовано правильно. Точність для тренувальних і тестових наборів становить 91,16% та 91,24% відповідно. Ці великі значення точності можуть бути пов'язані з дисбалансом класу у наборах даних.

Дисбаланс класу відбувається, коли відносна частота одного класу (наприклад, покупців, які придбали продукт) дуже низька в порівнянні з іншою (наприклад, клієнти, які не купували продукт). Це відбувається у багатьох сферах за різних причин, наприклад, у цифровому маркетингу (рейтинг кліків, коли лише невелика частина людей натискає посилання) чи у банківській справі (виявлення шахрайства, адже більшість трансакцій не шахрайські).

Проблеми зі точністю прогнозу виникають при незбалансованому наборі даних. Розглянемо це на прикладі email кампанії. Припустимо, що модель прогно-

зує не перехід користувач за надісланим посиланням. Тоді для тих користувачів, які все таки перейшли за посиланням модель завжди буде робити неправильний прогноз.

Оцінимо точність цієї моделі з незбалансованим набором даних (з 10000 клієнтів лише 100 переходів).

$$Accuracy = \frac{numberOfCorrectPreditions}{totalNumberPredictions} * 100\% = 99\%$$

Для цієї ж моделі оцінимо точність збалансованого набору даних (із 10 000 клієнтів, 5000 натискань).

$$Accuracy = \frac{5000}{10000} * 100\% = 50\% \cdot$$

Висока точність, що спостерігається з незбалансованим набором даних, пояснюється природою набору даних, а не продуктивністю моделі.

На жаль, дисбаланс класу негативно впливає на коректну роботу моделі, тобто при тренуванні моделі на незбалансованому наборі даних прогноз буде некоректним.

Висновки з проведеного дослідження. Для побудови біноміальної логістичної регресії було використано статистичні дані про характеристики потенційних споживачів продукту, а саме їх вік, стать, рівень освіти, сімейний стан, позицію на роботі тощо. Модель реалізована у середовищі R-Studio за допомогою вбудованого пакету caret. Отримана модель оцінена за допомогою показника загальної точності. Наведено можливі причини високих значень результату точності моделі. Якість моделі оцінено як високу, тому вона може бути рекомендована для застосування у банківському маркетингу.

Список використаних джерел:

1. Nehrey M., Hnot, T. Using recommendation approaches for ratings matrixes in online marketing. *Studia Ekonomiczne.* 2017. № 342. С. 115–130.

2. Негрей М.В., Гнот Т.В. Компаративний аналіз ефективності рекомендаційних систем в маркетингу. *Вісник Хмельницького національного університету. Економічні науки.* 2017. № 5. С. 278–286.

3. Elsalamony H.A. Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications.* 2014. № 85 (7). P. 12–22.

4. Sing'oei L., Wang J.. Data mining framework for direct marketing: A case study of bank marketing. *International Journal of Computer Science Issues (IJCSI).* 2013. № 10 (2 Part 2). P. 198.

5. Parlar T. Using data mining techniques for detecting the important features of the bank direct marketing data. *International journal of economics and financial issues.* 2017. № 7 (2). P. 692.

6. Wang D. Research on Bank Marketing Behavior Based on Machine Learning. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture.* 2020, October. Pp. 150–154.

7. Ghosh S., Hazra A., Choudhury B., Biswas P., Nag A. A comparative study to the bank market prediction. In *International Conference on Machine Learning and Data Mining in Pattern Recognition.* 2018, July. Springer, Cham. Pp. 259–268.

8. Wisaeng K.A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE).* 2013. № 3 (4). P. 116–119.

9. Grzonka D., Suchacka G., Borowik B. Application of selected supervised classification methods to bank marketing campaign. *Information Systems in Management.* 2016. № 5 (1). P. 36–48.

10. Yang S.B., Chen T.L. Uncertain decision tree for bank marketing classification. *Journal of Computational and Applied Mathematics.* 2020. № 371. P. 112710.

11. Yang T., Qian K., Lo D. C. T., Xie Y., Shi Y., Tao L. Improve the prediction accuracy of Naive Bayes classifier with association rule mining. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS).* 2016, April. Pp. 129–133.

12. Krishna C.L., Reddy P.V.S. Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques. *International Journal of Recent Technology and Engineering (IJRTE).* 2019. № 8 (3). P. 4373–4378.

13. Verma A. Evaluation of Classification Algorithms with Solutions to Class Imbalance Problem on Bank Marketing Dataset using WEKA. *International Research Journal of Engineering and Technology.* 2019. P. 54–60.

References:

1. Nehrey M., Hnot T. (2017) Using recommendation approaches for ratings matrixes in online marketing. *Studia Ekonomiczne,* no. 342, pp. 115–130.

2. Nehrey M., Hnot T. (2017) Komparatyvnyy analiz efektyvnosti rekomendatsiynykh system v marketynhu. *Visnyk Khmel'nyts'koho natsional'noho universytetu. Ekonomichni nauky,* no. 5, pp. 278–286.

3. Elsalamony H.A. (2014) Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications,* no. 85 (7), pp. 12–22.

4. Sing'oei L., Wang J. (2013) Data mining framework for direct marketing: A case study of bank marketing. *International Journal of Computer Science Issues (IJCSI)*, no. 10 (2 Part 2), p. 198.

5. Parlar T. (2017) Using data mining techniques for detecting the important features of the bank direct marketing data. *International journal of economics and financial issues,* no. 7 (2), p. 692.

6. Wang D. (2020, October) Research on Bank Marketing Behavior Based on Machine Learning. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture*, pp. 150–154.

7. Ghosh S., Hazra A., Choudhury B., Biswas P., Nag A. (2018, July) A comparative study to the bank market prediction. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 259–268). Springer, Cham.

8. Wisaeng K. (2013) A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE),* no. 3 (4), pp. 116–119.

9. Grzonka D., Suchacka G., Borowik B. (2016) Application of selected supervised classification methods to bank marketing campaign. *Information Systems in Management,* no. 5 (1), pp. 36–48.

10. Yang S.B., Chen T.L. (2020) Uncertain decision tree for bank marketing classification. *Journal of Computational and Applied Mathematics,* no. 371, p. 112710.

11. Yang T., Qian K., Lo D.C.T., Xie Y., Shi Y., Tao L. (2016, April) Improve the prediction accuracy of Naïve Bayes classifier with association rule mining. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 129–133). IEEE.

12. Krishna C.L., Reddy P.V.S. (2019, September) Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques. *International Journal of Recent Technology and Engineering (IJRTE),* no. 8 (3), pp. 4373–4378.

13. Verma A. (2019) Evaluation of Classification Algorithms with Solutions to Class Imbalance Problem on Bank Marketing Dataset using WEKA. *International Research Journal of Engineering and Technology,* pp. 54–60.

E-mail: Lzomchak@gmail.com