**УДК 519.254:519.237.8:616.831:616.13**

О. М. Мацуга[1], С. О. Дудукіна[2], С. П. Григорук[2]
*[1]Дніпровський національний університет імені Олеся Гончара*
*[2]КЗ «Дніпропетровська обласна клінічна лікарня імені І. І. Мечникова»*

# ПОБУДОВА МОДЕЛІ ПРОГНОЗУВАННЯ РЕЗУЛЬТАТУ ЛІКУВАННЯ НА ПРИКЛАДІ ОДНІЄЇ МЕДИЧНОЇ ЗАДАЧІ

**Розглянуто технологію побудови моделі прогнозування результату лікування у вигляді бінарної логістичної регресії. В рамках технології запропоновано процедуру відбору інформативних ознак на основі методу рекурсивного вилучення ознак. У ході практичної апробації технології побудовано дві моделі, що дозволяють прогнозувати результат реваскуляризації першого басейну при необхідності реваскуляризації обох у хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій при послідовній тактиці хірургічного лікування.**

**Ключові слова:** *модель прогнозування, бінарна логістична регресія, незбалансовані класи, відбір інформативних ознак, реваскуляризація судинного басейну, поєднані атеросклеротичні ураження церебральних та коронарних артерій.*

**Рассмотрена технология построения модели прогнозирования результата лечения в виде бинарной логистической регрессии. В рамках технологии предложена процедура отбора информативных признаков на основе метода рекурсивного исключения признаков. В процессе практической апробации технологии построены две модели, позволяющие прогнозировать результат реваскуляризации первого бассейна при необходимости реваскуляризации обоих у больных с сочетанными атеросклеротическими поражениями церебральных и коронарных артерий при последовательной тактике лечения.**

**Ключевые слова:** *модель прогнозирования, бинарная логистическая регрессия, несбалансированные классы, отбор информативных признаков, реваскуляризация сосудистого бассейна, сочетанные атеросклеротические поражения церебральных и коронарных артерий.*

**The paper considers the technology of building a treatment outcome prediction model in the form of binary logistic regression. The technology implies three stages. The first stage involves data preprocessing which includes anomaly detecting, missing values handling, one-hot encoding of categorical features, analyzing the correlations between features in order to exclude highly**

**correlated features and classes balancing. The second stage involves a feature selection. The feature selection procedure based on the recursive feature elimination method was suggested for this stage. The third stage implies learning of binary logistic regression on a dataset with the selected features. Newton's method was applied in order to learn the model in the research. To evaluate the model performance such metrics as accuracy, sensitivity and specificity were measured using 7-fold cross-validation technique. Two models for predicting the outcome of revascularization of the first basin while revascularization of the both basins is required in patients with combined atherosclerotic lesions of the cerebral and coronary arteries were built during practical testing of the technology. The first model predicts whether the outcome of revascularization will be favorable (1, 2, 3, 5) or not (4, 6) with 97% accuracy. If it predicts a favorable outcome, then the second model distinguishes the outcome 1, 5 from 2, 3 with 62% accuracy. The suggested feature selection procedure allowed significant reduction in the features number as well as building models without overfitting, whose performance was comparable to that of the model on the full set of features. The features number was reduced to 8 for the first model and 4 for the second model (initially there were 78 features). The scikit-learn python library and the Jupyter Notebook application were used to build the models.**

**Keywords:** *prediction model, binary logistic regression, imbalanced classes, feature selection, vascular basin revascularization, combined atherosclerotic lesions of the cerebral and coronary arteries.*

**Постановка проблеми.** Прогнозування результату лікування є актуальною проблемою і для хворих, і для лікарів, і для науковців. З позицій теорії машинного навчання побудова моделі прогнозування результату лікування є задачею класифікації, якщо результат лікування є категоріальною ознакою (наприклад, стан хворого поліпшився/не змінився/погіршився). Для розв'язання цієї задачі запропоновано достатньо багато моделей та підходів. Проте якість її розв'язання залежить не лише від вибору підходящої моделі, а й від технології підготовки та попередньої обробки даних, на яких будується модель. Як наслідок, процес побудови моделі часто суттєво залежить від даних і є потреба в його розробці для окремої практичної задачі.

**Мета** даної роботи полягала у побудові моделі прогнозування результату реваскуляризації одного басейну у хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій, яким показана реваскуляризація обох басейнів. Слід зазначити, що в медичній практиці питання послідовності проведення хірургічної реваскуляризації судинних басейнів у таких хворих є дискусійним. Тому прогнозування результату реваскуляризації першого басейну є вкрай важливе для вибору подальшої тактики лікування.

**Аналіз останніх досліджень і публікацій.** Існуючі моделі прогнозування результату реваскуляризації судинних басейнів [1–3] не враховують окремо періопераційні фактори ризику ускладнень та базуються тільки на загальних результатах лікування. У хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій, яким показано оперативне втручання на обох басейнах, моделей для прогнозування результату реваскуляризації басейну, що оперують першим, автори наразі не знають.

Оскільки побудова моделі прогнозування результату реваскуляризації є задачею класифікації, нижче наведено огляд основних моделей класифікації та обґрунтовано вибір однієї з них.

Широко вживаними на практиці є такі моделі: класифікатор *k* найближчих сусідів, дерево рішень, логістична регресія, машина опорних векторів, різновиди баєйсівського класифікатора, нейронні мережі та ансамблі класифікаторів (зокрема випадковий ліс та ансамбль на основі градієнтного бустінгу) [4]. Ансамблі класифікаторів вважають найбільш якісними моделями, проте їх результати важко інтерпретувати медичним фахівцям. Такий самий недолік мають і нейронні мережі, що фактично являють собою чорну скриньку. Крім того, для навчання нейронних мереж часто потрібні досить великі навчальні вибірки. Класифікатор *k* найближчих сусідів та різновиди байєсівського класифікатора не придатні для застосування, коли у наборі даних є одночасно кількісні та якісні ознаки, а у даній роботі задано саме такий набір даних. Дерево рішень – це досить проста для розуміння та інтерпретації модель, яку можна будувати на даних з різними типами ознак без особливої підготовки даних (як то масштабування кількісних ознак, бінаризація якісних ознак, опрацювання пропусків, відбір ознак тощо). Проте дерево рішень часто виявляється неоптимальним щодо якості класифікації. Недоліком машини опорних векторів у контексті поставленої практичної задачі можна вважати потребу в масштабуванні кількісних ознак, внаслідок чого певною мірою втрачається фізична інтерпретація моделі. З огляду на вищесказане, у роботі віддано перевагу логістичній регресії, зокрема бінарній логістичній регресії. Крім того, саме ця модель знайшла найбільш широке використання в медичній практиці.

Бінарна логістична регресія задається таким чином [4; 5]. Якщо для хворого можливі два результати лікування (А чи Б), тоді ймовірність результату А визначається як

$$\mathrm{P}\{\text{Результат} = \text{А}\} = \frac{1}{1+e^{-z}},$$

де

$$z = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p;$$

$a_0, a_1, a_2, \ldots, a_p$ – параметри моделі; $x_1, x_2, \ldots, x_p$ – значення ознак (симптомів) хворого.

Ймовірність результату Б, відповідно, дорівнює

$$\mathrm{P}\{\text{результат} = \text{Б}\} = 1 - \mathrm{P}\{\text{результат} = \text{А}\}.$$

Рішення щодо того, який саме результат має місце, приймають таким чином:

$$\text{Результат} = \begin{cases} \text{А}, & \text{якщо } \mathrm{P}\{\text{результат} = \text{А}\} > \text{поріг}, \\ \text{Б}, & \text{якщо } \mathrm{P}\{\text{результат} = \text{А}\} \le \text{поріг}. \end{cases}$$

Значення порога необхідно підбирати з огляду на бажані значення метрик якості.

Навчання бінарної логістичної регресії проводять за методом максимальної правдоподібності. Для знаходження максимуму функції правдоподібності використовують методи оптимізації, як правило, це різновиди методу градієнтного спуску або методу Ньютона [5].

Слід зазначити, що якість навченої моделі суттєвим чином залежить від результатів підготовки та попередньої обробки даних.

Залежно від специфіки даних підготовка та попередня обробка може включати в себе пошук аномалій, опрацювання пропущених значень, бінаризацію якісних та масштабування кількісних ознак, вилучення дублюючих ознак, балансування класів [6].

Важливим етапом підготовки даних також є відбір інформативних ознак. Існуючі методи відбору інформативних ознак можна розділити на три групи: фільтри, обгортки та вбудовані [6]. Методи фільтри, на відміну від обгорток, більш прості та швидкі, проте не враховують можливі зв'язки між ознаками. Вбудований відбір ознак притаманний лише деяким методам навчання. Наприклад, для логістичної регресії відбір ознак може мати місце у разі її навчання з $l_1$ регуляризацією [4].

Отже, модель прогнозування результату реваскуляризації басейну у хворих доцільно будувати у вигляді логістичної регресії. При цьому особливої уваги потребує підготовка та попередня обробка даних для навчання моделі.

**Постановка задачі.** В рамках дослідження, проведеного на базі комунального закладу «Дніпропетровська обласна клінічна лікарня

імені І. І. Мечникова», зібрано результати обстеження та лікування 195 хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій, яким показана хірургічна реваскуляризація обох судинних басейнів, проведено реваскуляризацію одного з басейнів. Результати обстеження та лікування представлено у вигляді матриці

$$\{x_{i,1}, x_{i,2}, \dots, x_{i,p}, y_i; i = \overline{1,N}\},$$

де $x_{i,j}, j = \overline{1,p}$ – значення $j$-ї ознаки (симптому), що описує стан $i$-го хворого на початку лікування; $y_i$ – результат реваскуляризації басейну для $i$-го хворого; $N$ = 195 – кількість хворих.

Серед ознак, які описують стан хворого на початку лікування, є кількісні, якісні категоріальні та якісні бінарні. Загальна кількість ознак дорівнює 78. У деяких хворих значення певних ознак відсутні.

Результат реваскуляризації басейну являє собою категоріальну ознаку, яка приймає одне з шести значень: 1 – покращення загального стану; 2 – повний регрес неврологічної/кардіологічної симптоматики; 3 – частковий регрес неврологічної симптоматики/зменшення класу стенокардії; 4 – погіршення стану (новий неврологічний дефіцит/збільшення класу стенокардії); 5 – без змін; 6 – смерть. Результати 1, 2, 3, 5 можна вважати сприятливими, а 4, 6 – несприятливими. При цьому хворих з несприятливими результатами у наборі даних трохи менше 3 % від загальної кількості.

На основі цих даних необхідно побудувати модель прогнозування результату реваскуляризації басейну.

**Основний матеріал.** Враховуючи, що результат реваскуляризації є категоріальною ознакою, у роботі мала місце задача класифікації з 6-ма класами. В ході проведених досліджень було встановлено, що результати 1 та 5 майже неможливо відрізнити один від одного, так само як і результати 2 та 3. Внаслідок цього було вирішено звести задачу багатокласової класифікації до двох бінарних задач. У ході розв'язання першої задачі будувалася модель для прогнозування того, буде результат реваскуляризації сприятливий (1, 2, 3, 5) чи несприятливий (4, 6). Під час розв'язання другої задачі будувалася модель для передбачення того, буде результат 1, 5 чи 2, 3. В обох випадках модель будувалася у вигляді бінарної логістичної регресії.

Побудову кожної бінарної логістичної регресії було здійснено за такою технологією:

1. Підготовка та попередня обробка даних. Враховуючи особливості набору даних, вона включала в себе:

– пошук аномалій; у заданому наборі їх не було виявлено;

– обробку пропущених значень; пропуски можна видаляти або заповнювати, наприклад, середнім чи медіанним значенням ознаки; у роботі під час побудови фінальних моделей проводилося заповнення пропусків середнім значенням;

– бінаризацію якісних категоріальних ознак;

– аналіз кореляційних зв'язків між ознаками з метою вилучення сильно корельованих ознак;

– балансування класів; у роботі віддано перевагу балансуванню шляхом дублювання випадково обраних представників малочисельного класу; при цьому потреба у балансуванні була лише під час побудови першої моделі.

2. Відбір інформативних ознак. У ході дослідження з метою скорочення простору ознак та зменшення перенавчання було здійснено спробу навчити бінарну логістичну регресію з регуляризацією. Проте за такого підходу не вдалося суттєво зменшити перенавчання. Тому було вирішено здійснювати відбір інформативних ознак за допомогою методів обгорток, серед яких обрано метод рекурсивного вилучення ознак (RFE), оскільки він єдиний реалізований у бібліотеці scikit-learn [7], що використовувалася в роботі. На основі методу RFE запропоновано таку процедуру. Спочатку здійснюється масштабування кількісних ознак. Далі у циклі M разів випадково обирається представницька частина з набору даних і для неї визначається T найважливіших ознак за допомогою методу RFE. Тоді для кожної ознаки підраховується, скільки разів вона була серед T найважливіших протягом M ітерацій (чим частіше, тим інформативніша ознака). Якщо для побудови моделі є потреба у балансуванні, то на кожній ітерації замість вибору представницької частини виконується випадковим чином балансування класів.

3. Навчання бінарної логістичної регресії на наборі з відібраними ознаками за методом максимальної правдоподібності. Для знаходження максимуму функції правдоподібності у роботі було використано метод Ньютона. Навчання проводилося без регуляризації. Значення порога для моделі визначалося із умови балансу чутливості та специфічності.

Для оцінки якості моделі було використано такі метрики як точність (accuracy), чутливість (sensitivity) та специфічність (specificity). Їх оцінювання було проведено за допомогою техніки 7-блокового ковзного контролю.

Під час практичної апробації описаної технології було побудовано

дві моделі, які дозволяють прогнозувати результат реваскуляризації першого басейну при необхідності реваскуляризації обох у хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій при послідовній тактиці хірургічного лікування. Побудову обох моделей виконано за допомогою python-бібліотеки scikit-learn [7] у додатку Jupyter Notebook.

Побудову першої моделі, яка призначена для передбачення несприятливого (4, 6) чи сприятливого (1, 2, 3, 5) результату реваскуляризації, було здійснено в умовах дуже сильної незбалансованості класів (менше 3% хворих мали несприятливі результати). Тому було виконано балансування класів шляхом дублювання випадково обраних представників малочисельного класу. Завдяки запропонованій процедурі відбору інформативних ознак було знайдено 8 ознак, на основі яких побудовано модель, якість якої співставна з якістю моделі на повному наборі ознак, при цьому не має перенавчання. Отже, кількість ознак було скорочено в 10 разів. Якість одержаної моделі виявилася досить високою, хоча модель потребує уточнення на основі набору даних, який би містив більшу кількість хворих з несприятливими результатами.

Ймовірність несприятливого результату за першою моделлю задається виразом:

$$\mathrm{P}\{\text{Результат} = \text{несприятливий}\} = \frac{1}{1+e^{-z}},$$

де

$$z = -51{,}51 - 17{,}326 \cdot \mathrm{I}(\text{Ускладнення після операції} = \text{немає}) -$$
$$-14{,}666 \cdot \mathrm{I}(\text{Ускладнення після операції} = \text{гострий інфаркт міокарду}) +$$
$$+16{,}183 \cdot \mathrm{I}(\text{Ускладнення після операції} = \text{стенокардія}) +$$
$$+7{,}597 \cdot \mathrm{I}(\text{Ускладнення після операції} = \text{ішемічний інсульт}) +$$
$$+25{,}485 \cdot \mathrm{I}(\text{Стать} = \text{чоловіча}) + 3{,}037 \cdot \mathrm{I}(\text{Рубцові зміни на ЕКГ} = \text{є}) +$$
$$+5{,}303 \cdot \mathrm{I}(\text{Транзиторна ішемічна атака} = \text{є}) +$$
$$+20{,}597 \cdot \mathrm{I}\left(\begin{array}{c}\text{Перетоки по передньосполучній артерії за}\\ \text{даними церебральної ангіографії} = \text{немає}\end{array}\right);$$

де $\mathrm{I}(\text{Умова})$ – індикаторна функція:

$$\mathrm{I}(\text{Умова}) = \begin{cases} 1, & \text{якщо Умова справедлива,} \\ 0, & \text{якщо Умова хибна.} \end{cases}$$

Рішення про результат реваскуляризації можна приймати за правилом:

$$\text{Результат} = \begin{cases} \text{несприятливий}, & \text{якщо } \mathrm{P}\{\text{Результат} = \text{несприятливий}\} > 0,5, \\ \text{сприятливий}, & \text{якщо } \mathrm{P}\{\text{Результат} = \text{несприятливий}\} \le 0,5. \end{cases}$$

Чутливість даної моделі (відсоток правильних прогнозів несприятливих результатів) дорівнює 75 %. Специфічність (відсоток правильних прогнозів сприятливих результатів) становить 97 %. Загальний відсоток правильних прогнозів (точність) дорівнює 97 %.

Якщо перша модель передбачає сприятливий результат, тоді за допомогою другої моделі можна відрізнити результати 1, 5 від 2, 3. Під час побудови другої моделі за допомогою запропонованої процедури відбору інформативних ознак вдалося знайти 4 ознаки, які забезпечили побудову моделі без перенавчання, якість якої не поступається якості моделі на повному наборі ознак. Тим самим кількість ознак було скорочено майже у 20 разів.

У другій моделі ймовірність результату 2, 3 задається виразом:

$$\mathrm{P}\{\text{Результат} = 2 \text{ або } 3\} = \frac{1}{1+e^{-z}},$$

де

$$z = 2,346 - 1,067 \cdot \mathrm{I}(\text{операція} = \text{послідовне СЦА} + \text{АКШ}) -$$

$$-0,037 \cdot \text{Вік} - 0,025 \cdot \text{Стеноз М1 сегменту середньомозкової артерії} +$$

$$+0,846 \cdot \mathrm{I}\begin{pmatrix} \text{Стеноз/звитість хребцевої артерії} \\ \text{за даними доплерографії} = \text{є} \end{pmatrix}.$$

Якщо рішення про результат реваскуляризації приймати з порогом 0,5, тобто на основі правила

$$\text{Результат} = \begin{cases} 2 \text{ або } 3, & \text{якщо } \mathrm{P}\{\text{Результат} = 2 \text{ або } 3\} > 0,5, \\ 1 \text{ або } 5, & \text{якщо } \mathrm{P}\{\text{Результат} = 2 \text{ або } 3\} \le 0,5, \end{cases}$$

тоді загальний відсоток правильних передбачень за другою моделлю складає 62 %. Чутливість моделі (відсоток правильних передбачень результатів 2, 3) дорівнює 70 %, а специфічність (відсоток правильних передбачень результатів 1, 5) – 55 %.

Якщо рішення приймати з порогом 0,53, тобто на основі правила

$$\text{Результат} = \begin{cases} 2 \text{ або } 3, & \text{якщо } \mathrm{P}\{\text{Результат} = 2 \text{ або } 3\} > 0,53, \\ 1 \text{ або } 5, & \text{якщо } \mathrm{P}\{\text{Результат} = 2 \text{ або } 3\} \le 0,53, \end{cases}$$

тоді точність, чутливість та специфічність моделі дорівнюють 62 %.

Розглянута у роботі технологія була також успішно апробована під час побудови моделі прогнозування наявності ускладнень у процесі реваскуляризації судинного басейну у тих самих хворих.

**Висновки.** У роботі розглянуто технологію побудови моделі прогнозування результату лікування у вигляді бінарної логістичної регресії.

1. В рамках технології запропоновано процедуру відбору інформативних ознак, яка базується на методі рекурсивного вилучення ознак.

2. У ході практичної апробації технології побудовано дві моделі, що дозволяють прогнозувати результат реваскуляризації першого басейну при необхідності реваскуляризації обох у хворих з поєднаними атеросклеротичними ураженнями церебральних та коронарних артерій при послідовній тактиці хірургічного лікування. Перша модель дозволяє передбачати, буде результат сприятливим чи ні, з точністю 97 %. Якщо вона передбачає сприятливий результат, тоді за допомогою другої моделі можна відрізнити результати 1, 5 від 2, 3 з точністю 62 %.

3. Запропонована процедура відбору інформативних ознак дозволила для першої моделі скоротити кількість ознак у 10 разів, а для другої моделі – в 20 разів.

## Бібліографічні посилання

1. Mohr F. W., Morice M. C., Kappetein A. P., Feldman T. E., Stahle E., Colombo A. et al. Coronary artery bypass graft surgery vs. percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. *Lancet*. 2013. 381 (9867). P. 629–638.

2. Соколова Н. Ю., Голухова Е. З. Реваскуляризация миокарда у больных стабильной ишемической болезнью сердца: стратификация периоперационных и отдаленных рисков. *Креативная кардиология*. 2016. 10 (1). С. 25–36.

3. Спирин Н. Н., Малышев Н. Н., Малышева И. В. Оценка и прогнозирование результатов каротидной эндартерэктомии клинико-математическим методом. *Журнал неврологии и психиатрии.* 2012. № 6. С. 40–44.

4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. *Data Mining, Inference, and Prediction*. 2009. 745 p.

5. Hosmer D., Lemeshow S. Applied Logistic Regression. 2nd ed. 2000. 375 p.

6. Kuhn M., Johnson K. Applied Predictive Modeling. Springer. 2013. 600 p.

7. Scikit-learn documentation. URL: https://scikit-learn.org/stable/index.html (дата звернення: 16.11.2020)

*Надійшла до редколегії 16.11.2020.*