3. Піддубна О.О. Економіко-математичне моделювання в управлінні виробничим потенціалом. *Економіка та держава*. 2009. № 12. С. 49-50.

УДК 519.2:004.8:004.62

Хмелівський Ю.С., студент СО «Магістр» спеціальності 122 «Комп'ютерні науки»
Нескородєва Т.В., к.т.н., доцент, зав. кафедри комп'ютерних наук та інформаційних технологій

АНАЛІЗ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ СЕРЦЕВОЇ НЕДОСТАТНОСТІ ЗАСОБАМИ МОВИ R

Донецький національний університет імені Василя Стуса, м. Вінниця

Серцева недостатність – це важкий стан, який виникає, коли серце не може нормально і достатньо «качати» кров по людському тілу. Це може бути гострим і раптовим захворюванням або прогресуючим, тривалим станом.[1]

Серцево-судинні захворювання (ССЗ) є причиною смерті номер 1 у всьому світі, щороку забирають приблизно 17,9 мільйона життів, що становить 31% усіх смертей у світі. Чотири з 5 смертей від серцево-судинних захворювань спричинені серцевими нападами та інсультами, і одна третина цих смертей трапляється передчасно у людей віком до 70 років. Серцева недостатність є поширеною подією, спричиненою ССЗ, і цей набір даних містить 11 ознак, які можна використовувати для прогнозування можливого захворювання серця.

Люди з серцево-судинними захворюваннями або люди з високим серцево-судинним ризиком (через наявність одного або кількох факторів ризику, таких як гіпертонія, діабет, гіперліпідемія або вже встановлене захворювання) потребують раннього виявлення та лікування, при цьому модель машинного навчання може бути дуже корисною.

Даний датасет створений для визначення захворювання серця.[2] Він містить 918 унікальних рядків та 11 стовпців із характеристиками людини і один бінарний стовпець відгук.

Атрибути:

1. Age: вік пацієнта [років]
2. Sex: стать пацієнта [M: Чоловік, Ж: Жінка]
3. ChestPainType: тип болю в грудях [TA: Типова стенокардія, ATA: Атипова стенокардія, NAP: неангінальний біль, ASY: безсимптомний]
4. RestingBP: артеріальний тиск у стані спокою [мм рт.ст.]
5. Cholesterol: сироватковий холестерин [мм/дл]
6. FastingBS: рівень цукру в крові натще [1: якщо FastingBS > 120 мг/дл, 0: інакше]
7. RestingECG: результати електрокардіограми в спокої [Normal: Нормальний, ST: наявність аномалій зубця ST-T (інверсії зубця T та/або елевація або депресія ST

> 0,05 мВ), LVH: ймовірна або визначена гіпертрофія лівого шлуночка за критеріями Естеса]

8. MaxHR: досягнута максимальна частота серцевих скорочень [числове значення від 60 до 202]
9. ExerciseAngina: стенокардія, спричинена фізичним навантаженням [Y: Так, N: Ні]
10. Oldpeak: oldpeak = ST [Числове значення, виміряне в депресії]
11. ST_Slope: нахил сегмента ST піку вправи на кардіограмі [Up: вгору, Flat: рівна, Down: вниз]
12. HeartDisease: вихідний клас [1: захворювання серця, 0: нормально].

В нас є 5 кількісних предиктора, тому потрібно їх перевести в числове значення.Sex та ExerciseAngina мають 2 рівня тому кожну з них можна одразу інтерпретувати як 0 та 1. Однак ChestPainType, RestingAngina та ST_Slope мають більше ніж 2 рівня, в такі ситуації ми можемо створити додаткові індикаторні змінні.[3]

Розглянемо цей механізм на прикладі змінної ChestPainType – тип болю в грудині. Створюємо три індикаторні змінні, перша з них виглядає так:

$$xi1 = \begin{cases} 1, \text{якщо тип болю } ASY \\ 0, \text{якщо тип болю не } ASY \end{cases}$$

друга має вигляд:

$$xi2 = \begin{cases} 1, \text{якщо тип болю } ATA \\ 0, \text{якщо тип болю не } ATA \end{cases}$$

і третя:

$$xi3 = \begin{cases} 1, \text{якщо тип болю } NAP \\ 0, \text{якщо тип болю } TA \end{cases}$$

Тоді всі ці змінні можна використовувати в рівнянні регресії для отримання наступної моделі:

$$yi = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + \beta 3 xi3 + \varepsilon i =$$
$$\begin{cases} \beta 0 + \beta 1 + \varepsilon i, \text{якщо тип болю } ASY \\ \beta 0 + \beta 2 + \varepsilon i, \text{якщо тип болю } ATA \\ \beta 0 + \beta 3 + \varepsilon i, \text{якщо тип болю } NAP \\ \beta 0 + \varepsilon i, \quad \text{якщо тип болю } TA \end{cases}$$

Розіб’ємо дані на навчальну та контрольну вибірки. В навчальну вибірку потраплять всі рядки з парним значенням предиктора Age, в тестову навпаки непарні. Також слід перевірити вибірки на репрезентативність.

Далі побудуємо логістичну модель з цілю передбачення HeartDisease на основі всіх предикторів. Данна модель показала частоту помилок в 15.45%. Розглянемо детальний опис властивостей підігнаної моделі.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.4846211  2.3591013  -1.477 0.139650
Age         -0.0082570  0.0203909  -0.405 0.685524
RestingBP    0.0088285  0.0095174   0.928 0.353611
Cholesterol -0.0044895  0.0016424  -2.734 0.006265 **
FastingBS    1.7424293  0.4284432   4.067 4.76e-05 ***
MaxHR       -0.0006226  0.0075211  -0.083 0.934023
Oldpeak      0.6292248  0.1809562   3.477 0.000507 ***
cptASY       2.9539925  0.6747785   4.378 1.20e-05 ***
cptATA       0.8935451  0.7246318   1.233 0.217538
cptNAP       0.6874976  0.6630992   1.037 0.299832
sex          1.7213539  0.4215795   4.083 4.44e-05 ***
ecgNormal   -0.0929554  0.3983734  -0.233 0.815500
ecgST       -0.0169567  0.5249726  -0.032 0.974233
eaYes        1.0595979  0.3664569   2.891 0.003834 **
stFlat       0.8745778  0.6970366   1.255 0.209585
stUp        -1.6338177  0.7224189  -2.262 0.023723 *
---
```

Рисунок 5 - Властивості логістичної регресії

В моделі є багато статистично незначимих предикторів, оскільки їх p статистика досить велика. Використання предикторів, не пов’язаних з відгуком, зазвичай збільшують частоту помилок на контрольній вибірці. Тому виділимо найбільш корисні для передбачення, і ми зможемо отримати більш ефективну модель. Це предиктори FastingBS, cptASY, sex та stUp.

На основі цих предикторів побудуємо логістичну регресію, а також LDA модель (лінійний дискримінантний аналіз), QDA модель (квадратичний дискримінантний аналіз) та KNN модель (метод K найближчих сусідів) з різним значенням K. В наступній таблиці можна розглянути частоту помилок для кожної моделі.

Модель	Частота помилки, %
Glm	14,34
LDA	14,34
QDA	16,11
KNN, K=1	16,77

KNN, K=8	14,56

Таблиця 1 - Частоти помилки моделей

Отже, найкращими моделями є логістична регресія та лінійний дискримінантний аналіз з помилками в 14,34%.

Визначимо коефіцієнти моделей та побудуємо їх.

Модель логістичної регресії:

$$y = 1.762 * FastingBS + 2.461 * cptASY + 1.889 * sex - 2.96 * stUp - 1.493$$

Модель лінійного дискримінантного аналізу:

$$y = 0.765 * FastingBS + 1.418 * cptASY + 0.881 * sex - 1.798 * stUp - 0.8622$$

Список літератури:

1. Серцева недостатність: симптоми та методи лікування [Електронний ресурс] . – Режим доступу до ресурсу: https://oxford-med.com.ua/ua/media-center/publikacii/serdechnaya-nedostatochnost/
2. Heart Failure Prediction Dataset [Електронний ресурс] . – Режим доступу до ресурсу: https://www.kaggle.com/fedesoriano/heart-failure-prediction
3. Джеймс Г., Уиттон А., Хасти Т., Тибширани Р. Введение в ститическое обучение с примерами на языке R Изд. Второе, испр. Пер с англ. С.Э. Мастицкого –М. ДМК Пресс, 2017. -456с

УДК 519.2:004.8:004.62

Чернега В.М., студентка 3 курсу спеціальності 122 «Комп'ютерні науки»
Нескородєва Т.В., к.т.н., доцент, доцент кафедри інформаційних технологій

АНАЛІЗ ФАКТОРІВ, ЩО ПРИЗВОДЯТЬ ДО СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ

Донецький національний університет імені Василя Стуса, м. Вінниця

На сьогоднішній день серцево-судинні захворювання – одна з найбільш частих причин смертності людей у світі. Вони пов'язані із патологіями в серцево-