УДК 519.24

Логистическая модель риска возникновения сердечно-сосудистых заболеваний

Сайфетдинов С.Ф., Егорова Д.К., Гарин М.А.

Национальный исследовательский Мордовский государственный университет

Аннотация: В настоящей работе рассматривается логистическая модель риска возникновения сердечно-сосудистых заболеваний. Модель реализована на языке Python с использованием библиотек в pandas и numpy. Приведена оценка модели.

Ключевые слова: набор данных, логистическая модель, вероятность, оценки качества, Python.

1. Набор данных

Для построения модели риска возникновения сердечно-сосудистых заболеваний был использован набор данных из открытого источника Kaggle [1]. Набор содержал количественные (пол, возраст, вес, рост, артериальное давление) и категориальные (никотиновая зависимость, употребление алкоголя, занятия спортом, наличие заболеваний сердечно-сосудистой системы) данные, собранные со слов 70000 пациентов на момент медицинского обследования.

Данные были очищены от аномальных значений с помощью вычисления верхнего и нижнего квантилей, нахождения их разницы с последующим вычислением нижней и верхней границ данных. Таким образом из 70000 строк осталось порядка 62000, что является достаточным для дальнейшего анализа.

2. Моделирование и оценка качества

Так как необходимо построить модель риска, по сути вероятность возникновения заболевания, то использовалось следующее регрессионное уравнение (логит-преобразование):

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_8 x_8)}},$$

здесь $x_1, x_2, ..., x_8$ — входные признаки приведенные выше, $\beta_0, \beta_1, \beta_2, ..., \beta_8$ — коэффициенты модели.

Очищенные данные разделили по полу пациентов, это обусловлено наличием существенных физиологических различий. Далее, полученные два набора, разделили на обучающую и тестовую выбрки и строили две модели для женщин и мужчин отдельно. Модели обучили с помощью метода `LogisticRegression` языка программирования `Python`.

Получили следующие оценки коэффициентов модели для пациентов-женщин:

$$\beta_0 = -11.4375, \beta_1 = 0.0499, \beta_2 = -0.0123, \beta_3 = 0.0178, \beta_4 = 0.0684, \beta_5 = 0.0133,$$
$$\beta_6 = -0.1327, \beta_7 = -0.2361, \beta_8 = -0.2587.$$

Для оценки качества данной модели использовались следующие метрики:

· доля правильных ответов модели (`Accuracy` = 0.7137),

· доля объектов, которые действительно принадлежат классу относительно всех объектов, которые модель отнесла к этому классу (`Precision` = 0.7478),

· доля истинно положительных классификаций (`Recall` = 0.6569),

· F-мера, которая является гармоническим средним между `Precision` и `Recall` (`F-мера` = 0.6994),

· `ROC`-кривая – доля ложно положительных примеров в сравнении с долей правильно положительных примеров (`area`=0.78).

Оценка качества модели показала удовлетворительный результат, учитывая, что для ее построения использовалось 8 признаков.

Подобная модель реализована авторами в Android-приложении `«HLS»` [2].

Литература

1. Cardiovascular Disease dataset. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/discussion?sort=undefined

2. Свидетельство о государственной регистрации программы для ЭВМ № 2024663282. Программа-помощник контроля здорового образа жизни «HLS» : заявл. 30.05.2024 : опубл. 05.06.2024 / Т. И. Власова, Д. К. Егорова, Д. В. Пузакова, М. А. Гарин, С. Ф. Сайфетдинов; правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва».

MSC 62J12

Logistic model of cardiovascular disease risk

S.F. Sayfetdinov, D.K. Egorova, M.A. Garin

National Research Mordovia State University

Abstract: The article presents a logistic model of the risk of cardiovascular diseases. The model is implemented in Python using libraries in pandas and numpy. The model estimate is given.

Keywords: data set, logistic model, probability, quality assessments, Python.

References

1. Cardiovascular Disease dataset. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/discussion?sort=undefined
2. Certificate of state registration of the computer program No. 2024663282. Healthy lifestyle control assistant program "HLS": application. 05/30/2024: publ. 06/05/2024 / T. I. Vlasova, D. K. Egorova, D. V. Puzakova, M. A. Garin, S. F. Sayfetdinov; copyright holder Federal State Budgetary Educational Institution of Higher Education «National Research Ogarev Mordovia State University».