АНАЛИЗ РИСКА АРТЕРИАЛЬНОЙ ГИПЕРТЕНЗИИ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

А. В. Хлебович

Белорусский государственный университет, г.Минск;
al.khlebovich@gmail.com
науч. рук. – В. И. Малюгин, канд. физ.-мат. наук, доц.

Целью работы является применение и сравнительный анализ эффективности алгоритмов статистического и машинного обучения в задачах оценки риска артериальной гипертензии. В работе на основе обследования одной и той же выборки пациентов с интервалом 5 лет построена бустинговая модель бинарной классификации, которая применена для оценки эффективности методики лечения в режиме продолженного наблюдения за пациентами. Исследовано влияние динамики основных факторов риска пациента на его медицинский диагноз и результаты статистической классификации.

Ключевые слова: риск артериальной гипертензии, факторы риска, продолженное наблюдение, алгоритмы машинного обучения.

ВВЕДЕНИЕ

В настоящее время все большее и большее распространение получает превентивная медицина. Превентивная медицина – это направление в современной медицинской науке и практике, главной целью которой является сохранение здоровья человека за счет предупреждения развития различных заболеваний и патологий [1].

При оценке риска артериальной гипертензии (АГ) актуальными являются две задачи [2]: оценка риска первичной артериальной гипертензии (ПАГ) и оценка риска развития АГ в режиме продолженного наблюдения в течение определенного периода времени. Для решения этих задач в работе используются алгоритмы бинарной классификации выборки пациентов на классы, которые в соответствии с медицинским диагнозом относятся к «здоровым» и «больным». При наличии классифицированной обучающей выборки данная задача может решаться с помощью различных алгоритмов статистического и машинного обучения. В данной работе для классификации пациентов в заданном пространстве признаков (факторов риска) использованы следующие алгоритмы: логистическая регрессия (Logistic Regression), дерево решений (Decision Tree), метод опорных векторов (SVM), бустинговые алгоритмы (XGBoost, LightGBM, CatBoost) [3-5].

ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ДАННЫХ

Данные для исследований предоставлены РНПЦ «Кардиология» Республики Беларусь и включают выборки результатов первичного и повторного обследования пациентов в режиме продолженного наблюдения через 5 лет. Начальная выборка наблюдений состоит из результатов обследования 589 пациентов. Выборка в соответствии с медицинским диагнозом разбита на два класса: класс здоровых (197 чел.) и класс больных пациентов, в который вошли пациенты с 1, 2 и 3 степенью заболевания ПАГ (392 человек). Повторная выборка наблюдений состоит из результатов обследования для 285 пациентов. При повторном обследовании класс здоровых включает 95 человек, и класс больных – 190 человек.

В качестве классификационных признаков используются такие факторы риска, как: пол (gender), возраст (age), индекс массы тела (bmi), абдоминальное ожирение или обхват талии в сантиметрах (ao), уровень физической активности (activity), курение (smoking), наличие семейного анамнеза ранних сердечно-сосудистых заболеваний со стороны родственников мужского пола (male_heredity).

ОЦЕНИВАНИЕ РИСКА ПАГ И ЗНАЧИМОСТИ ФАКТОРОВ РИСКА НА ЭТАПЕ ПЕРВИЧНОГО ОБСЛЕДОВАНИЯ

С использованием начальной выборки пациентов проведены эксперименты, целью которых являлось построение различных статистических моделей зависимости бинарной переменной, характеризующей степень риска, от факторов риска. Для оценки эффективности алгоритмов используемая выборка разбивается на обучающую выборку (train sample) объема 412 наблюдений и экзаменационную выборку (test sample), включающую 177 наблюдений.

Целью данного исследования является сравнение различных классификационных моделей и поиск той, которая обладает лучшей эффективностью по следующим характеристикам качества (метрикам): Precision (точность), Recall (полнота), ROC-AUC- мера для обучающей (train) и тестовой (test) выборок, F1-мера [3, 4].

Результаты экспериментов, представленные в сводной таблице 1, говорят о том, что алгоритм SVM с полиномиальным ядром проигрывает по всем параметрам. Для дальнейшего исследования выбираются алгоритм логистической регрессии, метод опорных векторов с радиальной базисной функцией (RBF) и бустинговые алгоритмы.

Таблица 1

Результаты построенных моделей

	Порог отсечения	ROC-AUC (train)	ROC-AUC (test)	F1-мера	Recall	Precision
Logistic Regression	0.75	0.80	0.73	0.80	0.83	0.77
Decision Tree	0.70	0.83	0.70	0.80	0.81	0.79
SVM (полиномиальное ядро)	0.65	0.72	0.62	0.76	0.86	0.69
SVM (RBF-ядро)	0.62	0.82	0.74	0.81	0.85	0.78
XGBoost	0.46	0.82	0.71	0.81	0.86	0.78
LightGBM	0.55	0.83	0.70	0.80	0.83	0.78
CatBoost	0.50	0.83	0.72	0.80	0.84	0.77

Для оценки важности признаков, соответствующих основным факторам риска АГ, используется алгоритм CatBoost. Согласно рисунку 1, полученному с помощью данного алгоритма, самыми значимыми факторами риска являются индекс массы тела и абдоминальное ожирение, наименьшая значимость у параметра пол, но он также имеет значение.

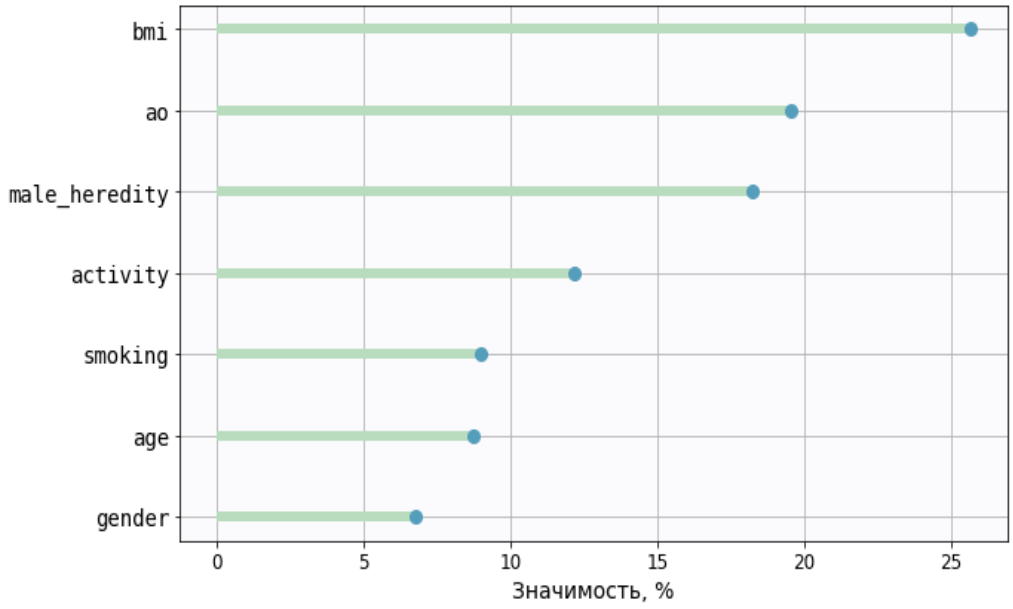

Рис. 1. Вклад признаков в предсказание риска заболевания

ОЦЕНКА ЭФФЕКТИВНОСТИ МОДЕЛЕЙ В РЕЖИМЕ ПРОДОЛЖЕННОГО НАБЛЮДЕНИЯ ЗА ПАЦИЕНТАМИ

Те же алгоритмы применялись для оценки состояния пациентов на повторном приёме спустя 5 лет. Тестирование проводилось на выборке из 285 пациентов. Все алгоритмы имели высокие показатели точности при классификации больных пациентов, поскольку никто из них не улучшил свое состояние. С классификацией здоровых по медицинскому диагнозу пациентов лучше всего справился бустинговый алгоритм LightGBM, его характеристики точности приведены в таблице 2.

Таблица 2

Точность алгоритма LightGBM в режиме продолженного наблюдения

Порог	ROC-AUC	F1-мера	Recall	Precision
0.55	0.75	0.8	0.83	0.77

Расхождение оценок состояния пациентов при повторном обследовании, полученных на основании полного медицинского обследования и статистической классификации на основе факторов риска, объясняется тем, что практически по всем факторам риска, используемым в классификационных алгоритмах, наблюдалось ожидаемое ухудшение состояния пациентов, связанное с возрастными изменениями и не соблюдением рекомендаций по лечению. Лучше всего, как и на первоначальном обследовании, показал себя алгоритм LightGBM, который рекомендован к дальнейшему применению.

Библиографические ссылки

1. Что такое превентивная медицина? // HBP[Electronic resource]- 2020. – Mode of access: https://hbp-group.ru/chto-takoe-preventivnaya-medicina/ – Date of access: 11.05.2020.
2. Pavlova O.S., Malugin,V.I. Computer Analysis of Essential Hypertension Risk on the Base of Genetic and Environmental Factors / O.S. Pavlova [et al.] // Proc. of the 11th Intern. Conf. «Computer Data Analysis and Modeling», Minsk. – 2016. – P. 289-293.
3. LightGBM Documentation // LightGBM [Electronic resource]- 2020. – Mode of access: https://lightgbm.readthedocs.io/en/latest/ – Date of access: 10.05.2020.
4. Scikit-learn. Machine learning in Python // Scikit-learn [Electronic resource]- 2020. – Mode of access: https://scikit-learn.org/stable/ – Date of access: 10.05.2020.
5. Aurélien Géron Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition/Aurélien Géron – O'Reilly Media, Inc., 2020. – 690 c.
6. XGBoost Documentation // XGBoost [Electronic resource]- 2020. – Mode of access: https://xgboost.readthedocs.io/en/latest/ – Date of access: 10.05.2020.