ВАЖНОСТЬ ПРЕДИКТОРОВ ПАЦИЕНТОВ С КЛЕЩЕВЫМИ ИНФЕКЦИЯМИ

В.С. Сафронов, магистрант гр. 8ПМ0И1,
Е.В. Сафронова, аспирант гр. А0-36
С.В. Аксёнов, к.т.н., доц. ОИТ ИШИТР
Томский политехнический университет
E-mail: vss75@tpu.ru

Введение

Исследования клещевых инфекций берут своё начало с непосредственного открытия инфекций, переносчиками которых являются клещи. В России первые описания клещевого энцефалита (КЭ) были сделаны в 1894 г. [1]. Болезнь Лайма (иксодовый клещевой боррелиоз, ИКБ) впервые описана в 1975 году как локальная вспышка артритов в г. Лайм (США) [2]. Ежегодно в регионах, климат которых благоприятен для существования данных насекомых, фиксируются случаи заражения. Несмотря на то, что данные заболевания известны человечеству более полувека и для каждого отдельного вируса и инфекции разработана методика лечения, некоторые случаи являются непредсказуемыми. Тяжелая форма заболевания может привести к инвалидизации и летальному исходу. В настоящее время благодаря развитию и широкому распространению информационных технологий, помимо хорошо известных специалистам в сфере медицины методов диагностики и контроля заболеваний, применяются инструменты аналитики данных и машинного обучения, что позволяет находить между теми или иными признаками пациентов неизвестные ранее зависимости, закономерности. Цель исследования – определение степени важности предикторов пациентов с клещевыми инфекциями.

Основная часть

Набор данных содержит информацию о 193 пациентах с диагнозами КЭ, ИКБ и микст. В результате чистки данных из более 150 было отобрано 97 признаков, которые характеризуют антропометрические характеристики пациентов, показатели анализов крови, информация о наличии тех или иных отклонений в состоянии здоровья, наличии сопутствующих заболеваний и т.д. Для обучения моделей классификации набор данных предварительно был разделен на обучающую и тестовую выборки в соотношении 0,7 к 0,3, соответственно. Для классификации использовались такие ансамблевые методы, основанные на деревьях решений, как случайный лес и градиентный бустинг. Эффективность работы обученных моделей была оценена с помощью таких метрик качества, как чувствительность и специфичность. Под чувствительностью понимают долю положительных результатов, которые правильно идентифицированы как таковые. Специфичность представляет собой долю отрицательных результатов, которые правильно идентифицированы как таковые. Учитывая тот факт, что классификация производится по трём диагнозам, то метрики рассчитываются для каждого из них. В качестве положительных классов выступает каждый конкретный диагноз, а отрицательные классы – это количество пациентов с другими двумя диагнозами. В таблице 1 указаны средние значения рассчитанных метрик по всем трем классам.

Таблица 1. Метрики качества работы моделей классификации

Алгоритм классификации	Чувствительность	Специфичность
Случайный лес	0,804	0,788
Градиентный бустинг	0,803	0,799

В 80,4% и в 80,3% случаев пациенты верно отнесены к определенному диагнозу и в 78,8% и в 79,9% случаев верно определено, что к данному диагнозу пациенты не относятся, случайным лесом и градиентным бустингом, соответственно. Помимо оценки качества обучения моделей также было проведено ранжирование предикторов по степени важности влияния на результат классификации. Для этого применялся такой метод интерпретации моделей машинного обучения, как SHAP (Аддитивные объяснения Шепли). Аппроксимация векторов Шепли, служащих для определения оптимального распределения выигрыша между игроками в теории игр, является результатом оценки важности предикторов методом SHAP [3]. На рисунках 1 и 2 представлена визуализация важности предикторов по отношению к диагнозу согласно SHAP значениям для моделей случайный лес и градиентный

бустинг, соответственно. В список наиболее важных признаков вошли жалобы: Светобоязнь, Тошнота, Заторможенность, Сыпь, Слабость, Боли в мышцах, Головная боль, Боли в суставах; отклонения в состоянии: Ригидность затылочных мышц, Наличие эритемы на месте укуса, Покраснение миндалин, Симптом Кернига, Поза Ромберга; анализы крови: BAS%, LYM%, NEU%, RBC, HGB, EOS%, WBC.

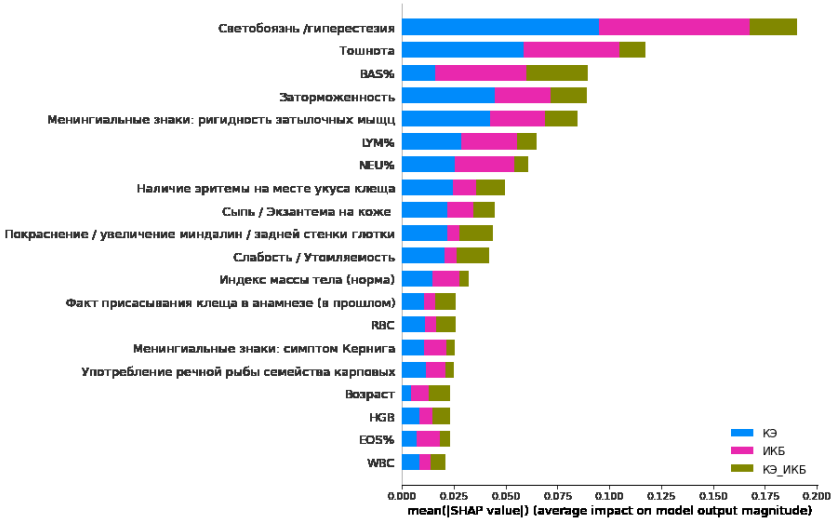

Рис. 1. Степень важности признаков модели «Случайный лес»

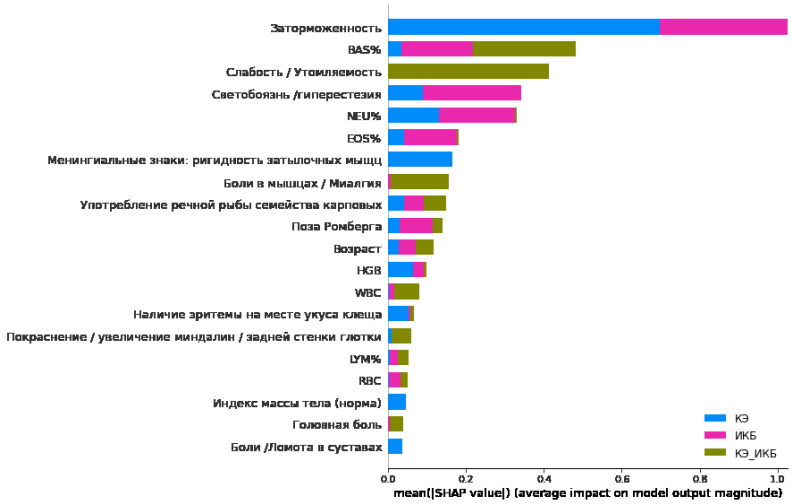

Рис. 2. Степень важности признаков модели «Градиентный бустинг»

Заключение

В результате проведения исследования можно сделать вывод о том, что используемые модели классификации показали приблизительно одинаковые результаты. Чувствительность случайного леса выше на 0,1%, а специфичность градиентного бустинга выше на 1,1%. Что касается важности предикторов, такие показатели, как Светобоязнь, Заторможенность, содержания в крови базофилов (BAS%) являются наиболее важными для обеих моделей.

Список использованных источников

1. Осторожно – клещевой энцефалит! // Министерство здравоохранения Хабаровского края [Электронный ресурс]. – URL: https://zdrav.khv.gov.ru/node/178 (20.01.2022).
2. Диагностика, лечение и профилактика клещевого энцефалита и иксодового клещевого боррелиоза у военнослужащих МО РФ: методические указания / составители: К. В. Жданов [и др.] – Москва: МО РФ, 2018. – 62 с.
3. Интерпретируй это: метод SHAP в Data Science // Чернобровов Алексей – аналитик [Электронный ресурс]. – URL: https://chernobrovov.ru/articles/interpretiruj-eto-metod-shap-v-data-science.html (10.02.2022).