РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
ФИЗИКО-ТЕХНИЧЕСКИЕ ПРОБЛЕМЫ РАЗРАБОТКИ ПОЛЕЗНЫХ ИСКОПАЕМЫХ

2024 № 2

ГОРНАЯ ИНФОРМАТИКА

УДК 53.05 + 539.3

МОДЕЛЬ ПРОГНОЗА ПРИРОДНОГО ПОЛЯ НАПРЯЖЕНИЙ В ПЛОТНОМ ПЕСЧАНИКЕ НА ОСНОВЕ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ XGBOOST

Ду Тун, Ли Юйвэй

Ляонинский университет,
E-mail: liyuweibox@126.com, *110036, г. Шэньян, Китай*

Алгоритм машинного обучения XGBoost используется для оценки природного поля напряжений. С применением метода корреляции Пирсона установлено, что характерными параметрами каротажа, наилучшим образом коррелирующими с минимальным значением горизонтального (тектонического) напряжения, являются данные спектрального гамма-каротажа, глубокого каротажа, индукционного каротажа, акустического каротажа, глубина залегания и содержание в породе кальция, а с максимальным значением горизонтального (тектонического) напряжения — глубина, данные спектрального гамма-каротажа, каротажа самопроизвольной поляризации, кавернометрии и плотностного каротажа. Результаты модели XGBoost сравнивались с моделью линейной регрессии, моделью опорных векторов и моделью случайного леса. Для проверки общей способности модели выполнена *k*-блочная перекрестная валидация. Показано, что алгоритм XGBoost позволяет прогнозировать природные напряжения в породе на основе малого объема исходных данных с точностью 94 % и высоким уровнем генерализации данных. Модель линейной регрессии обладает наибольшей скоростью расчета и минимальной точностью прогнозирования. Модели опорных векторов и случайного леса показали приемлемую точность. Полученные с помощью алгоритма XGBoost результаты универсальны и могут использоваться при решении проблем, связанных с прогнозированием природного поля напряжений в горных породах.

Природное поле напряжений, алгоритм XGBoost, плотный песчаник, машинное обучение

DOI: 10.15372/FTPRPI20240215
EDN: ZYRCVE

Природные напряжения — это естественные напряжения в земной коре, не вызванные инженерной деятельностью и изменяющиеся в течение продолжительного геологического времени под влиянием тектонических сдвижений. Их изменение влияет на разработку нефтегазовых месторождений и безопасность ведения подземных работ. Ввиду низкой пористости и низкой проницаемости нетрадиционных нефтегазовых коллекторов, коммерческая отдача

Работа выполнена в рамках программы по поддержке молодых исследователей Ляонинского университета (№ XLYC2203100).

таких пластов достигается благодаря крупномасштабной интенсификации добычи с помощью гидроразрыва. Основной фактор, влияющий на его эффективность, — местные (природные) напряжения. При разработке подземных месторождений такие напряжения — определяющий фактор разрушения пород вокруг выработок. Точное прогнозирование поля напряжений важно при разработке нетрадиционных нефтегазовых месторождений и для предотвращения горных ударов при подземных горных работах.

В настоящее время предложено множество методов прямого и косвенного измерения напряжений [1]. Методы прямого измерения включают микросейсмические методы, гидроразрыв, метод обрушения стенки скважины, акустическое измерение на основе эффекта Кайзера, геологическое картирование и т. д. [2 – 9], методы косвенного измерения — метод восстановления напряжений, анализ напряжений путем термоупругого бурения, анализ геологических структур, анализ неупругой деформации и др. [10 – 16]. Измерительные методы позволяют получать данные о напряжениях в породе, однако на точность данных влияет точность измерительного оборудования, к тому же такие измерения весьма затратны.

Для снижения затрат предложен ряд моделей расчета природного поля напряжений. В [14] разработана расчетная модель напряженного состояния одиночного разлома, залегающего близко к поверхности на основе его смещений, в [17] — модель расчета напряжений на основе критерия Хука – Брауна. Также существуют модели Мэттьюза и Келли, Андерсона, Кулона – Мора, Хуана, Гурра, комбинированная пружинная и др. [18 – 20].

В связи с последними достижениями в области регистрации данных некоторые исследователи применяли технические средства для прогнозирования напряжений в массиве. В [21] для их анализа использовалась фотосъемка. В [22, 23] регистрировались многополярные акустические сигналы с целью выявления трещин в породных формациях и определения ориентации природных напряжений для определения механических свойств породы. В [24] рассчитан коэффициент сплошности породного массива с помощью регистрации акустических сигналов, а также напряжения *in situ* на основе использования статических механических свойств. В [25] анализировались изображения кернов и петрофизические данные для изучения распределения напряжений в складчатом поясе Келасу.

Использовались также методы численного моделирования для анализа природных напряжений. В [26] описано текущее напряженное состояние газоносного района Dibei на основе каротажа скважин и численного моделирования. В [27] проведено исследование с использованием программы 3DECs для моделирования напряженного состояния карьеров Раваччоне и Фантискритти в бассейне Каррары (Италия). В [28] посредством программного обеспечения (ПО) UDEC и FLAC моделировалось влияние обратных разломов на развитие напряжений. В [29] с помощью метода конечных элементов построена геологическая модель угольных месторождений и смоделировано распределение напряжений в разные геологические периоды. В [30] предложены методы расчета двух- и трехмерных полей напряжений на основе нескольких измерительных точек и граничных условий на дневной поверхности. В [31] метод дискретных элементов в ПО 3DEC применялся для анализа напряженного состояния трещиноватого породного массива.

Перечисленные модели имеют свои недостатки. Большинство моделей по прогнозу напряжений в массиве — эмпирические, обладают региональными ограничениями, а также содержат большое количество параметров, которые необходимо рассчитывать по другим методам и формулам, что делает определение природных напряжений весьма громоздким. К тому же определение большого количества параметров увеличивает ошибки вычисления. При этом работы, рассматривающие природные напряжения, сфокусированы на традиционные месторождения, в то время как традиционные и нетрадиционные песчаные коллекторы существенно различаются по литологическим свойствам.

В связи с достижениями в области компьютерных технологий и непрерывной оптимизацией расчетных алгоритмов для прогнозирования напряжений стали применяться методы машинного обучения, анализирующие сложные нелинейные зависимости между большими наборами данных. Расчет природного поля напряжений на основе каротажных данных обладает уникальными преимуществами, так как каротаж осуществляется на значительной глубине и обеспечивает большой объем выборки. В [32] метод опорных векторов использовался для прогнозирования напряжений угля и породы на основе кавернометрии, компенсированного нейтронного каротажа, спектрального гамма-каротажа, плотностного каротажа и глубокой резистивиметрии. Рассмотренные данные обладают корреляцией с природными напряжениями и позволяют их непосредственный прогноз с точностью 94 %. В [33] нейронная сеть с обратным распространением применялась для прогнозирования механических свойств породы на основе данных акустического каротажа, а также выполнялось косвенное прогнозирование природного поля напряжений в сланце. Модели, в основе которых лежит метод опорных векторов или нейронная сеть с обратным распределением, обладают высокой точностью в условиях большого объема исходных данных, что в свою очередь при недостатке данных приводит к недо- или переобучению.

Недостаток данных ограничивает применение методов машинного обучения для прогнозирования напряжений в горной породе [34]. В настоящей работе предложено прямое прогнозирование напряжений в плотном песчанике на основе алгоритма XGBoost. Он преобразует формулу Тейлора второго порядка для функции потерь, так как альтернативная функция учитывает ограничения при прогнозировании. Определены характерные параметры модели по коэффициенту корреляции Пирсона с помощью каротажных данных из нескольких источников и построены следующие модели: XGBoost; линейной регрессии; случайного леса; опорных векторов. Все модели проверялись методом *k*-блочной перекрестной валидации на основе ряда оценочных критериев. Для прогнозирования природных напряжений на разной глубине участка D1 коллектора из плотного песчаника на нефтяном месторождении Daqing в Китае выбрана наиболее эффективная модель.

МЕТОДЫ ИССЛЕДОВАНИЯ

В области разработки нетрадиционных нефтегазовых месторождений и подземных инженерных работ природные напряжения можно разделить на вертикальное (гравитационное) напряжение, а также минимальное и максимальное горизонтальные (тектонические) напряжения. Гравитационная компонента напряжений возникает в результате давления налегающих пород, изменяется от плотности и глубины залегания формации h и рассчитывается на основе данных плотностного каротажа:

$$\sigma_v = \int_0^H \rho(h) g d h, \tag{1}$$

где $\rho(h)$ — функция изменения плотности формации при изменении h, г/см^3; g — ускорение свободного падения, м/с^2.

Если плотность измеряется с определенным интервалом по глубине, то данные плотности дискретны, в этом случае для расчета гравитационной компоненты напряжений аппроксимирующий интеграл плотности заменяется на кумулятивную сумму плотностей в каждой точке измерения:

$$\sigma_v = \rho_{\text{ave}} g h + \sum_i \rho_i g R_{LEV},$$

где ρ_{ave} — средняя плотность над целевым пластом; $R_{LEV} = 0.125$ м — интервал измерения плотности; ρ_i — плотность коллектора в i-м интервале.

Тектонические напряжения имеют две главные компоненты: напряжения в направлении x и y. Они вычисляются как

$$\sigma_{x1} = \frac{\mu}{1-\mu}(\sigma_v - \alpha p_p) + \alpha p_p, \tag{2}$$

$$\sigma_{y1} = \frac{\mu}{1-\mu}(\sigma_v - \alpha p_p) + \alpha p_p. \tag{3}$$

Здесь μ — коэффициент Пуассона; p_p — поровое давление породной формации, МПа; α — коэффициент эффективного напряжения (коэффициент Био).

Тектоническое напряжение — составляющая совокупного напряжения. Оно возникает в результате динамического поведения земной коры, т. е. сдвижения геологических структур или сейсмической активности и имеет очевидный анизотропный характер. Фактическое напряжение анизотропно, и вектор максимального горизонтального напряжения обычно сонаправлен с суммарным вектором тектонических напряжений. Главная составляющая тектонического напряжения — напряжение тектонических сдвижений в направлении x и y:

$$\sigma_{x2} = \xi_x(\sigma_v - \alpha p_p) + \alpha p_p,$$

$$\sigma_{y2} = \xi_y(\sigma_v - \alpha p_p) + \alpha p_p,$$

где ξ_x, ξ_y — коэффициенты тектонического напряжения (коэффициенты бокового отпора) в направлении x и y.

Для расчета максимального и минимального главных горизонтальных напряжений необходимо определить другие параметры: коэффициент Пуассона; константу Био; коэффициент бокового отпора и т. д. Добавление дополнительных параметров увеличивает объем измерений и повышает погрешность вычисления конечного значения, поэтому предлагается использовать методы машинного обучения для прямого прогнозирования горизонтального напряжения на основе данных каротажа скважин.

Рассматриваемый целевой коллектор расположен на участке D1 нефтяного месторождения Daqing и состоит из плотного песчаника. На участке D1 ранее получены каротажные данные (рис. 1), а именно: данные спектрального гамма-каротажа (GR), самопроизвольной поляризации (SP), кавернометрии (CAL), глубокого (LLD), поверхностного (LLS), микробокового (MSFL), акустического (AC), глубокого индукционного (ILD), среднего индукционного (ILM) каротажа, а также глубина (D), плотность (DEN), содержание песчаника (SAND), глины (VSH) и кальция (VCA). Так как данные разных видов каротажа отражают соотношения между физическими параметрами коллектора, они представляют собой большой объем информации, относительно непрерывны и могут использоваться для прогнозирования природного поля напряжений.

Исследование разделено на следующие этапы: определение характерных параметров с помощью коэффициента корреляции Пирсона; нормализация данных; построение моделей машинного обучения (XGBoost, линейная регрессия (LR), случайный лес (RF), метод опорных векторов (SVM)) для выявления нелинейных зависимостей между характерными параметрами и природными напряжениями; оценка эффективности моделей и их сравнение между собой (рис. 1).

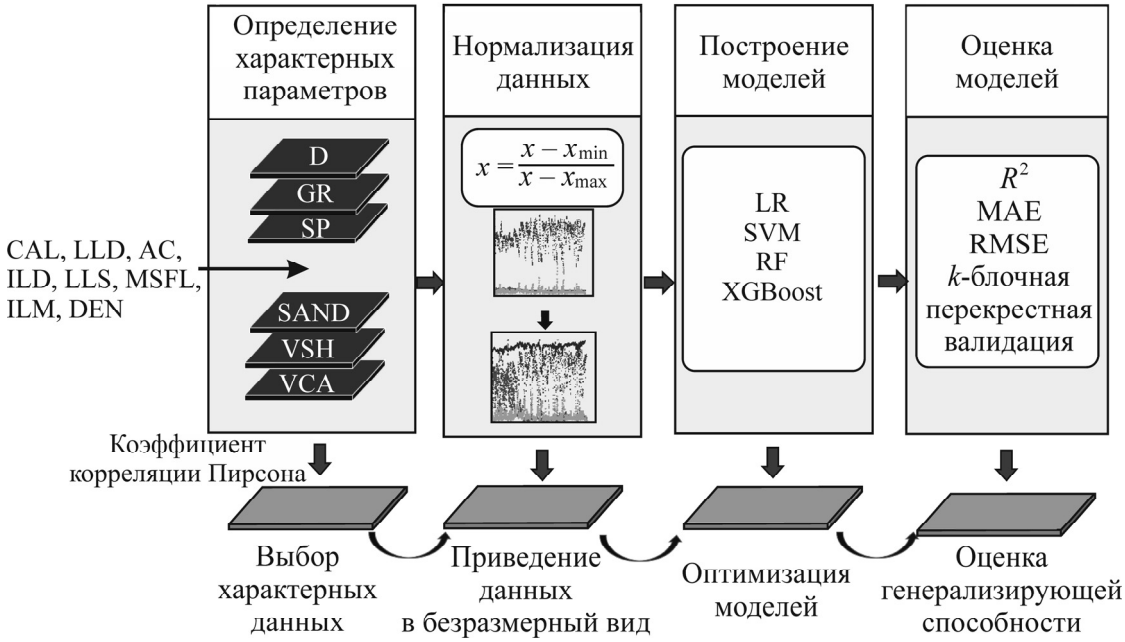

Рис. 1. Схема прогнозирования природного поля напряжений в нефтяном коллекторе из плотного песчаника месторождения Daqing на основе каротажных данных (их описание дано в тексте)

Определение характерных параметров. В совокупности рассмотрен массив из 1600 наборов данных каротажа скважины *T* на участке D1 коллектора из плотного песчаника нефтяного месторождении Daqing (рис. 2).

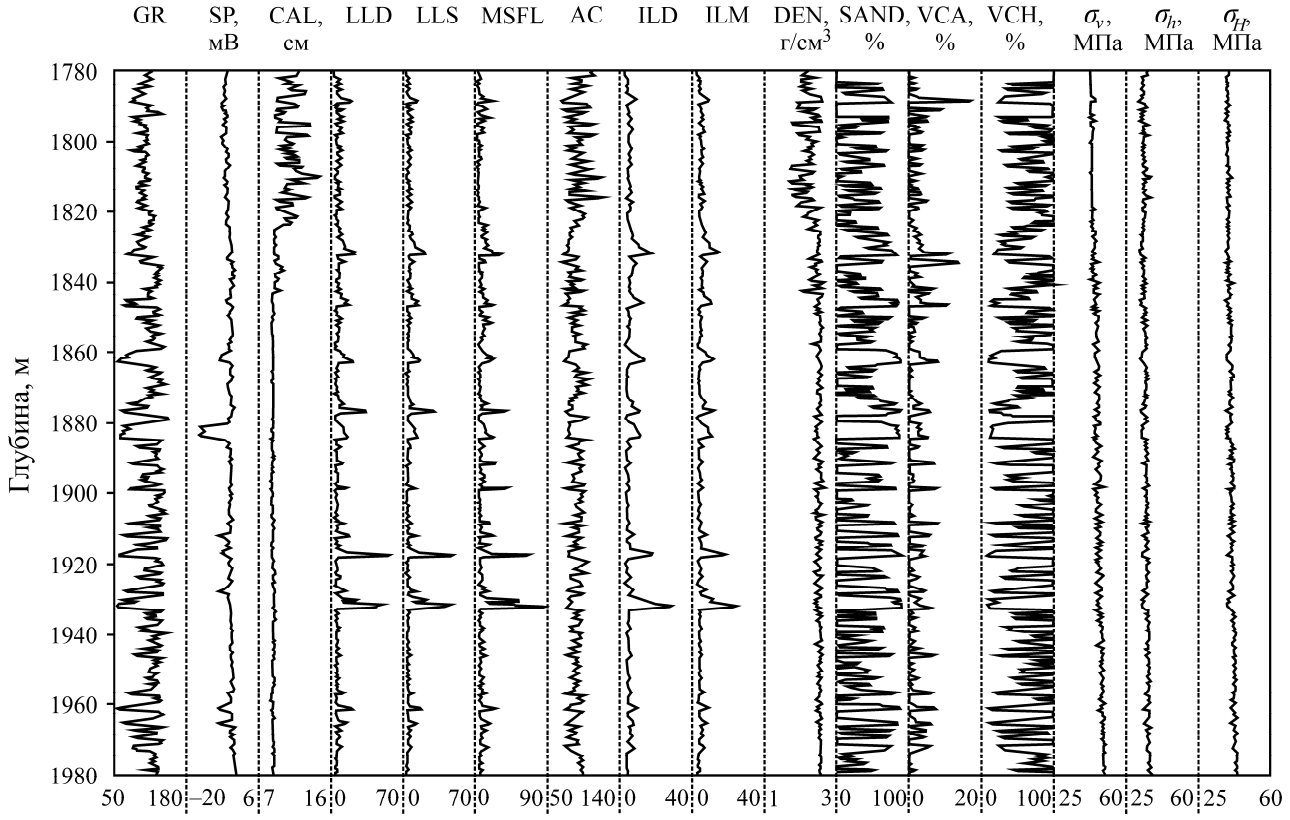

Рис. 2. Данные каротажа скважины *T* на участке D1

Для определения соответствия данных каротажа с напряжениями использовался коэффициент корреляции Пирсона τ, представляющий собой отношение отклонения от среднего и среднего квадратичного отклонения между двумя переменными:

$$\tau = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right).$$

Здесь $\overline{X}$ — среднее значение выборки X; X_i — i-е значение в выборке X; σ_X — среднее квадратичное отклонение выборки X; n — общее количество значений в выборке. При $\tau \to -1$ между переменными существует сильная отрицательная корреляция, при $\tau \to 1$ — сильная положительная корреляция, при $\tau \to 0$ — корреляция отсутствует.

На рис. 3 приведены результаты определения коэффициента Пирсона.

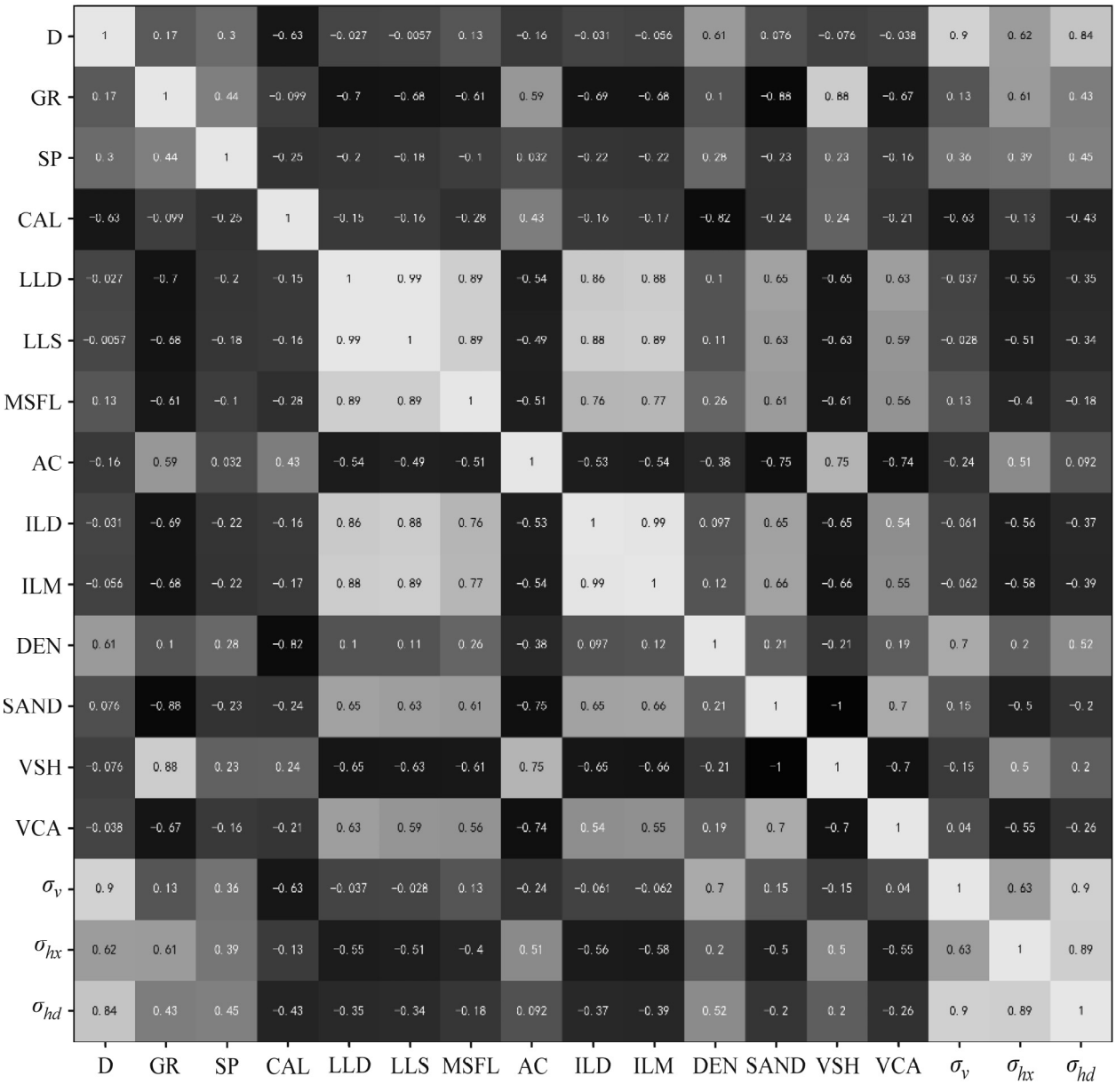

Рис. 3. Результаты расчета коэффициента корреляции Пирсона

Вертикальное (гравитационное) напряжение имеет наибольшую корреляцию с глубиной залегания и плотностью породы (параметры превышают 0.7). Этот результат соответствует формуле (1). Наибольшая корреляция выявлена между минимальным главным горизонтальным напряжением и глубиной (0.62). Эти величины имеют между собой сложную нелинейную зависимость. Далее наблюдается сильная корреляция между интенсивностью радиоактивности пород (GR) и минимальным главным горизонтальным напряжением (0.61). Обычно, чем больше содержание мадстоуна в породе, тем выше GR, так как GR отражает изменения литологического состава в коллекторе и напряжений. Сильная отрицательная корреляция отмечается между минимальным главным горизонтальным напряжением и LLD, LLS, ILD, ILM. Эти параметры отражают различия в электропроводности породы, показывая литологический состав массива и коэффициент Пуассона. В формулах (2), (3) коэффициент Пуассона — важный параметр, влияющий на напряжения; AC и VCA также влияют на минимальное главное горизонтальное напряжение.

Для минимального главного горизонтального напряжения введен пограничный коэффициент корреляции Пирсона, равный 0.5. Параметры, превышающие 0.5, считаются характерными. Коэффициенты параметров LLD и LLS, ILD и ILM близки к 0.5, поэтому в качестве характерных рассмотрены только LLD и ILD. Выявлены характерные параметры, наиболее влияющие на минимальное главное горизонтальное напряжение: D; GR; LLD; ILD; AC; VCA. Подобным образом проанализировано максимальное главное горизонтальное напряжение с пограничным значением 0.4 и выявлены следующие характерные параметры: D; DEN; GR; SP; CAL.

Нормализация данных. Параметры каротажа весьма не равномерны при использовании размерных величин, например глубина скважины может измеряться в тысячах метров, тогда как GR — в десятках. Требуется нормализация данных, дающая два преимущества: во-первых, она исключает негативное влияние размерности, упрощает вычислительную сложность модели и снижает или исключает необходимость обходных расчетов; во-вторых, сокращает трудности, связанные с низкой конвергенцией данных, продолжительным временем обучения и низкой точностью модели из-за слишком большого разброса данных.

В работе использовался дисперсионный метод стандартизации данных, позволяющий привести все характерные параметры в интервал [0, 1]. Для каждого параметра минимальное значение приводилось к 0, максимальное — к 1 по формуле $x_i^{norm} = (x_i - x_{\min}) / (x_{\max} - x_{\min})$, где x_i^{norm} — i-е нормализованное значение характерного параметра; x_i — i-е значение характерного параметра x; $x_{\max}$, $x_{\min}$ — минимальное и максимальное значения характерного параметра x. На рис. 4 представлены результаты нормализации характерных параметров.

Модель линейной регрессии. Линейная регрессия — простейший регрессионный алгоритм:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n ,$$

где θ_0 — точка пересечения; $\theta_1 \sim \theta_n$ — коэффициенты. Данное выражение можно представить в виде матрицы:

$$z = [\theta_0, \theta_1, \theta_2 ... \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} = \theta^T x (x_0 = 1) .$$

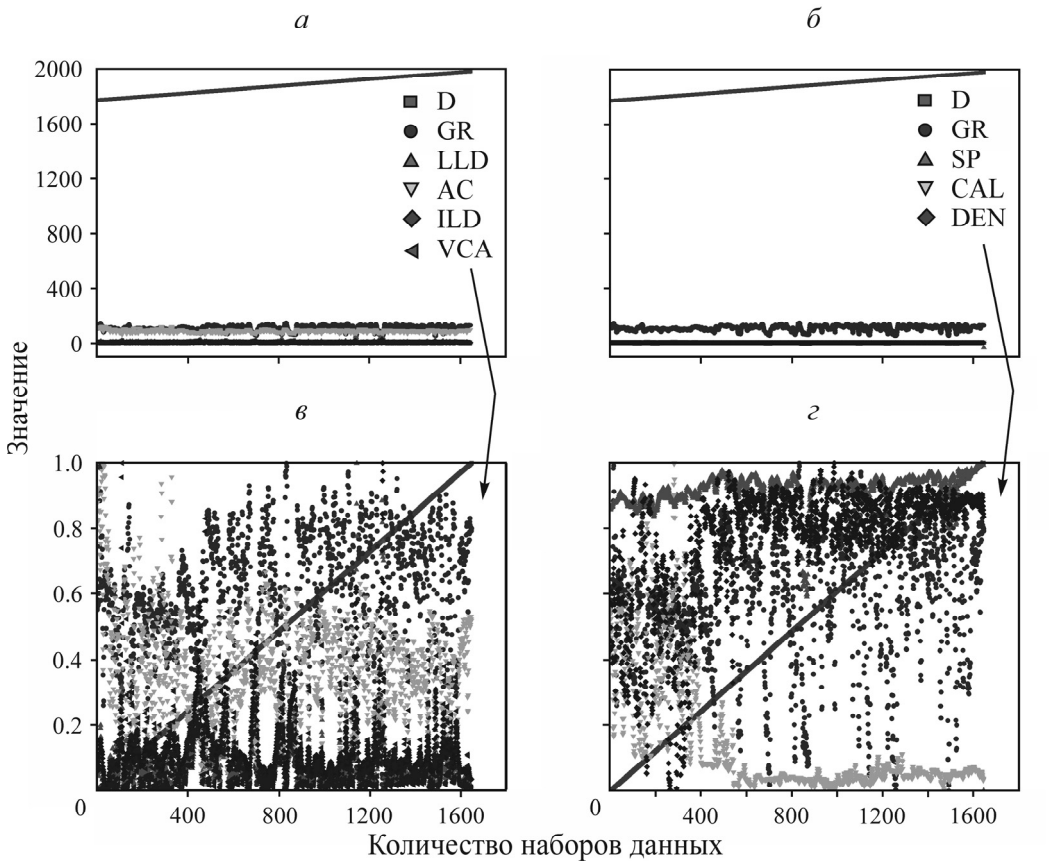

Рис. 4. Распределение характерных параметров до нормализации данных (*а*, *б*) и после нее (*в*, *г*): *а*, *в* — минимальное главное горизонтальное напряжение; *б*, *г* — максимальное

Основная задача линейной регрессии заключается в построении прогнозирующей функции z для выявления линейной зависимости между входной характерной матрицей x и меткой значений y. Основа построения прогнозирующей модели — определение θ^T и θ_0 методом наименьших квадратов (рис. 5). Используя функцию z, на основе матрицы характерных параметров x модель линейной регрессии предлагает ряд значений меток $y_{\text{спрог}}$ для решения различных задач по прогнозированию непрерывных величин.

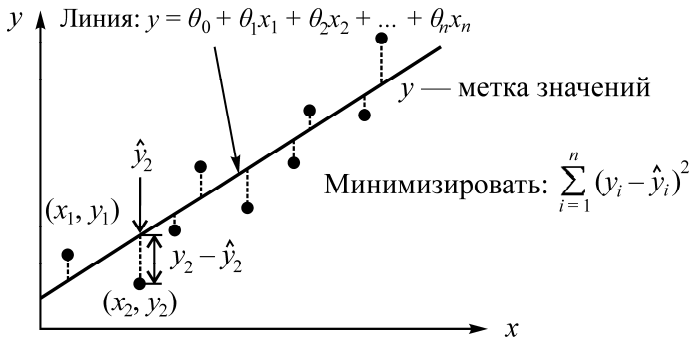

Рис. 5. Линейная регрессия по методу наименьших квадратов

Преимущества алгоритма линейной регрессии:

— прост и не требует большого количества времени для вычислений;

— высокая точность определения линейных зависимостей в наборах данных;

— эффективен при малом объеме входных данных;
— результаты предрасположены к интерпретации;
— не требуется корректировка параметров.

Тем не менее данный алгоритм не подходит для нелинейных данных, так как обладает низкой точностью и склонен к переобучению.

Модель опорных векторов. В ее основе лежит теория статистики, основная идея — определение гиперплоскости (границы принятия решений) для максимального уменьшения ошибки прогнозирования, особенно ошибки классификации малого массива исходных данных (рис. 6). Преимущества данного метода: обработка высокоразмерных данных; высокая генерализирующая способность; подходит для небольших выборок; решение нелинейных задач; хорошая надежность и способность к интерпретации данных.

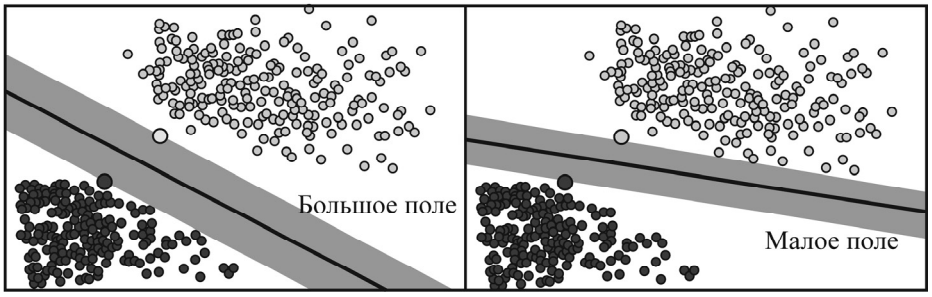

Рис. 6. Диаграмма метода опорных векторов

Модель случайного леса. Данный метод — оптимизированный алгоритм ансамблевого обучения, в рамках которого идея ансамблевого обучения интегрирована во множественные деревья решений (рис. 7). Метод выявляет наиболее надежные результаты среди большого количества базовых деревьев. Окончательный результат прогнозирования учитывает все модели деревьев. Для повышения точности и предотвращения переобучения от одиночного дерева происходит случайная обработка разных деревьев решений. Каждое дерево решения — независимый объект.

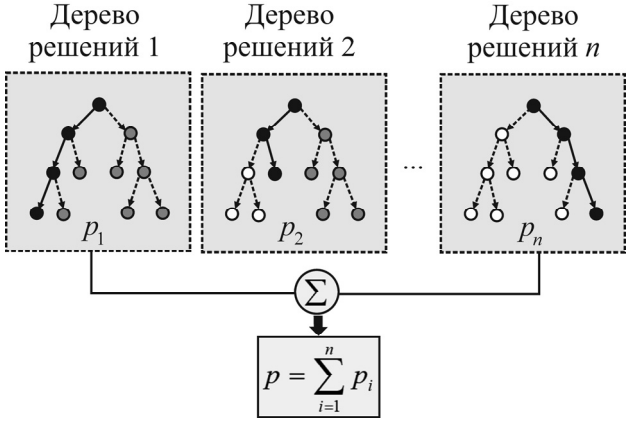

Рис. 7. Принцип метода случайного леса

Метод случайного леса позволяет:
— обрабатывать высокоразмерные данные без выбора характерных параметров;
— после обучения выявлять наиболее значимые характерные параметры;
— выполнять параллельный расчет, ускоряя вычислительный процесс;
— визуализировать процесс, упрощая последующий анализ результатов.

В выборках данных с большим уровнем шума данный метод склонен к переобучению. Характерные параметры с большим разделением значений существенно влияют на принятие решения, тем самым и на результат прогнозирования.

Модель XGBoost. Алгоритм XGBoost — форма критического градиентного бустинга. Градиентный бустинг обучает множество моделей постепенно, аддитивно и последовательно. Алгоритм позволяет преодолеть деревьям решений собственные вычислительные ограничения и достичь инженерных целей с высокой скоростью и эффективностью. Так как данный алгоритм относится к ансамблевым методам обучения, базой для обучения является массив деревьев решений, а спрогнозированное значение каждого дерева добавляется в окончательное значение. В алгоритме XGBoost используется формула Тейлора второго порядка для функции потерь в качестве суррогатной функции, позволяющей выявить оптимальную точку деления и узел выходного значения на дереве регрессии. XGBoost вводит значения подузлов и количество поддеревьев в функцию потерь, позволяя избежать переобучения, а также значительно повышает эффективность моделирования по сравнению с обычным градиентным бустингом деревьев решений (GBDT) за счет оценки точки разветвления регрессионного дерева и параллелизации подузлов совместно с характеристиками конвергентности второго порядка.

Преимущества алгоритма XGBoost:

— использует для предыдущей итерации формулу Тейлора второго порядка для функции потерь, тогда как GBDT — формулу первого порядка, поэтому XGBoost обладает бóльшей точностью и осуществляет такое же обучение за меньшее количество итераций;

— использует многопоточный выбор наиболее оптимальной точки сегментации, повышая скорость вычислений;

— вводит слагаемое регуляризации в функцию потерь для контроля за сложностью модели и для уменьшения вероятности переобучения.

Критерий оценки. Для оценки прогнозирования использовался коэффициент корреляции R^2, определяющий отклонение между двумя переменными величинами:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2},$$

где n — количество выборок; y_i — значение i-й точки данных; $\hat{y}_i$, $\overline{y}_i$ — спрогнозированное и среднее значения; числитель — сумма квадрата разностей между фактическим и спрогнозированным значениями; знаменатель — сумма квадрата разностей между фактическим и средним значениями, которая описывает степень дисперсии данных.

Коэффициент корреляции R^2 изменяется в интервале [0, 1]: 0 — отсутствие соответствия между спрогнозированным и фактическим значениями; 1 — отсутствие ошибок моделирования (полное соответствие). Также выполнялась оценка по средней абсолютной (MAE) и среднеквадратической (RMSE) ошибке:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|(y_i - \hat{y}_i)\right|,$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ .$$

Способность моделей машинного обучения к генерализации — важный показатель их эффективности. В процессе машинного обучения модели определяется функция потерь, которая минимизируется для повышения эффективности модели. Обучение модели выполняется для решения практических проблем, и простая минимизация функции потерь не гарантирует высокую эффективность или даже возможность решения рассматриваемых потерь.

Данная проблема решается k-блочной перекрестной валидацией — распространенным методом проверки генерализирующей способности моделей машинного обучения. Основная идея заключается в разделении массива данных на k отдельных блоков, где один блок проверочный, а $k-1$ блоки — обучающие. Процесс повторяется k-итераций, где каждый раз выбирается другой проверочный блок. В качестве индекса эффективности, по которому можно сделать вывод о способности модели генерализировать неизвестные данные, используется среднее значение k результатов валидации. На рис. 8 метод перекрестной валидации показан на примере разделения массива на 10 блоков.

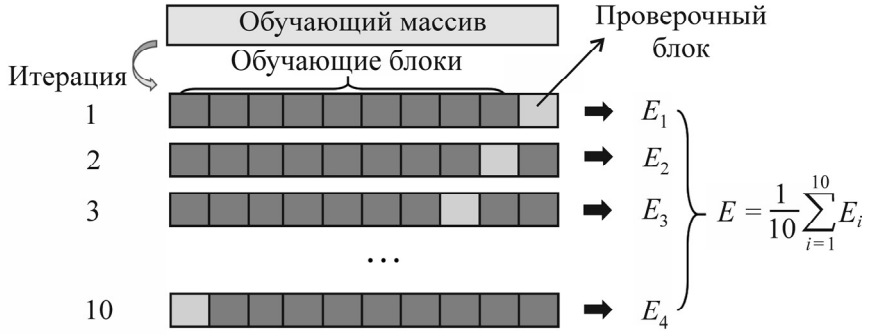

Рис. 8. Перекрестная валидация на примере разделения массива на 10 блоков

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Выполнено прогнозирование максимального и минимального главных горизонтальных напряжений с помощью четырех моделей машинного обучения. Проанализировано 1600 наборов данных, которые были разделены на обучающий и проверочный наборы в соотношении 8 : 2.

Из таблицы видно, что модель линейной регрессии имеет наибольшую скорость вычисления максимального главного горизонтального напряжения (среднее время расчета 0.00101 с), модель случайного леса — минимальную (1.46031 с); время расчета модели опорных векторов и XGBoost составило 0.05337 и 0.16002 с соответственно. Модель линейной регрессии выявляет только линейные зависимости между наборами данных, поэтому ее вычислительная скорость максимальна. Для модели случайного леса требуется обработать большое количество деревьев решений, так как окончательный прогноз учитывает все деревья, поэтому вычислительная скорость минимальна. Такой же характер скорости вычисления наблюдается и при прогнозировании минимального главного горизонтального вычисления: для модели линейной регрессии среднее время расчета составило 0.00099 с, для модели случайного леса — 1.76879 с. При прогнозе максимального главного напряжения наибольшую точность имеет модель XGBoost со средними значениями $R^2 = 97.5\,\%$, MAE = 0.185, RMSE = 0.062, что на 15.8, 4.1 и 1.2 % выше, чем точность модели линейной регрессии, опорных векторов и случайного леса. Аналогичный характер точности имеет прогнозирование минимального главного горизонтального напряжения. Для модели XGBoost средние значения R^2, MAE и RMSE составили 94.4 %, 0.248 и 0.108 соответственно.

Оценка эффективности рассматриваемых моделей

Главное горизонтальное напряжение	Модель	Время расчета, с	MAE	R^2	RMSE
		max / min / среднее	max / min / среднее	max / min / среднее	max / min / среднее
Максимальное	LR	0.00001 / 0.00200 / 0.00101	0.441 / 0.600 / 0.493	0.797 / 0.868 / 0.842	0.308 / 0.554 / 0.386
	RF	1.4282 / 1.5467 / 1.4603	0.180 / 0.236 / 0.212	0.958 / 0.977 / 0.963	0.057 / 0.109 / 0.088
	SVM	0.0285 / 0.0686 / 0.0534	0.265 / 0.326 / 0.288	0.920 / 0.953 / 0.937	0.119 / 0.209 / 0.152
	XGBoost	0.1521 / 0.1998 / 0.1600	0.157 / 0.204 / 0.185	0.966 / 0.979 / 0.975	0.053 / 0.075 / 0.062
Минимальное	LR	0.00099 / 0.00199 / 0.00099	0.376 / 0.512 / 0.436	0.788 / 0.878 / 0.838	0.239 / 0.452 / 0.313
	RF	1.7157 / 1.8204 / 1.7688	0.234 / 0.282 / 0.253	0.916 / 0.947 / 0.941	0.106 / 0.148 / 0.116
	SVM	0.0284 / 0.0805 / 0.0542	0.274 / 0.341 / 0.313	0.871 / 0.933 / 0.8965	0.135 / 0.265 / 0.179
	XGBoost	0.1662 / 0.1752 / 0.1689	0.221 / 0.271 / 0.248	0.921 / 0.958 / 0.944	0.078 / 0.141 / 0.108

Примечание: LR — линейная регрессия; RF — модель случайного леса; SVM — модель опорных векторов.

На рис. 9 приведены результаты перекрестной валидации.

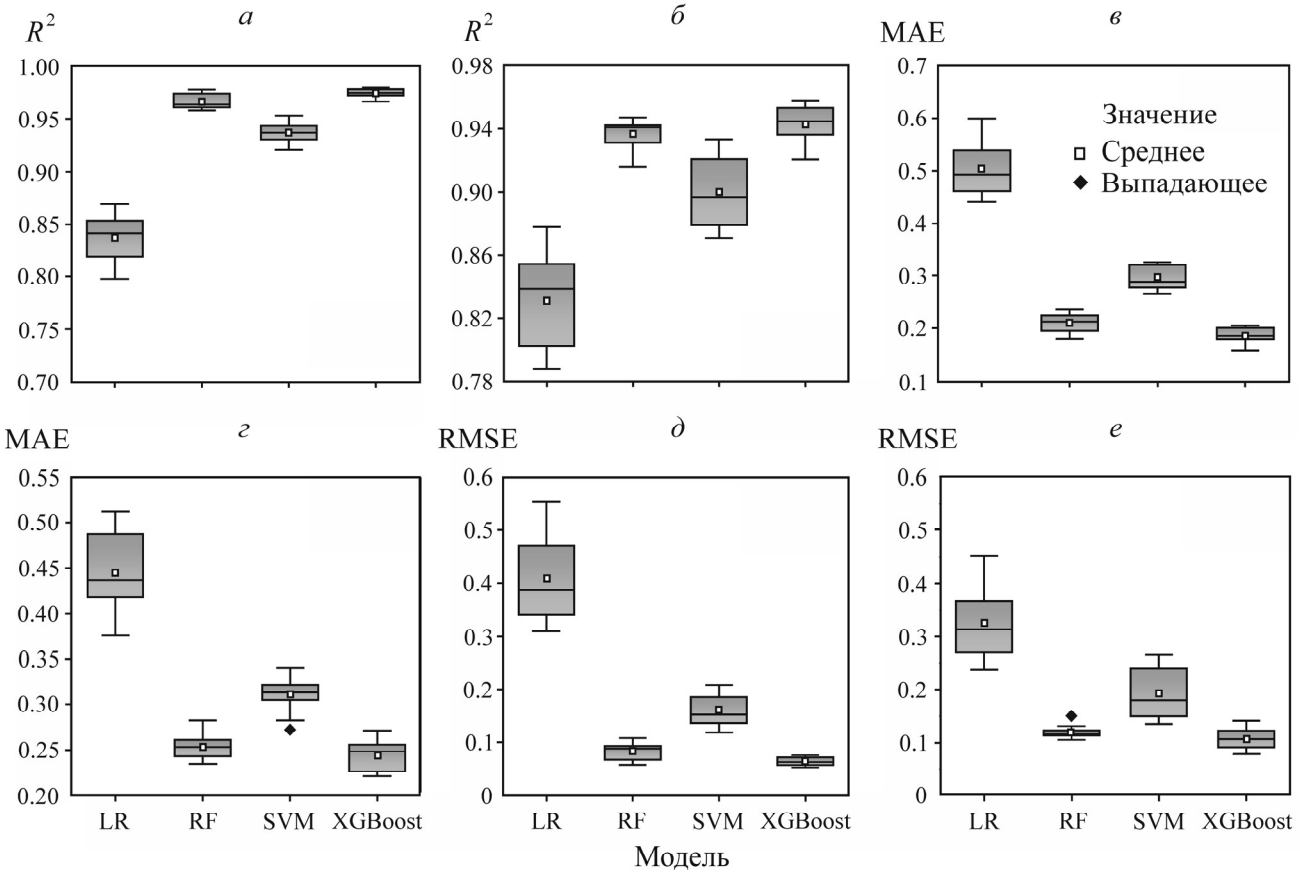

Рис. 9. Прямоугольная диаграмма R^2, MAE, RMSE: *а*, *в*, *д* — максимальные напряжения, *б*, *г*, *е* — минимальные

После 10 расчетных итераций результаты R^2 модели XGBoost более сконцентрированы, площадь участка наименьшая. Коэффициенты R^2 модели XGBoost наибольшие среди всех построенных. Точность *k*-блоковой перекрестной валидации относительно устойчива, модель

обладает высокой способностью генерализировать данные. Эффективность модели при обучении и проверке не различается, поэтому она способна обрабатывать неизвестные данные. Результаты 10 итераций модели линейной регрессии наиболее рассеяны (участок на рис. 9 имеет наибольшую площадь), R^2 — наименьшее, модель линейной регрессии имеет низкую устойчивость и точность при прогнозировании поля напряжений. Участок модели опорных векторов имеет относительно большую площадь, так как после регрессии метод опорных векторов не способен прогнозировать за пределами тренировочного массива данных, что приводит к переобучению при моделировании рассеянных данных. Если модель не проверена после обучения и переобучена, она может эффективно прогнозировать в рамках текущего массива данных, но не способна генерализировать данные в более широком диапазоне. Для XGBoost значения MAE и RMSE наименьшие. При прогнозировании максимального главного горизонтального напряжения результаты перекрестной валидации для XGBoost максимально сконцентрированы, а соответствующий прямоугольник имеет наименьшую площадь, демонстрируя наибольшую эффективность модели. При прогнозировании минимального главного горизонтального напряжения результаты перекрестной валидации для модели случайного леса максимально сконцентрированы и соответствующий прямоугольник имеет наименьшую площадь, показывая наибольшую эффективность модели.

Обученные модели применялись к тестовому массиву данных, состоящему из 20 % от общего массива. На рис. 10 показано изменение максимального и минимального главных горизонтальных напряжений для четырех моделей в зависимости от глубины. Сплошная линия — фактическое напряжение $y_{\text{факт}}$, кружок — спрогнозированные значения $y_{\text{прог}}$.

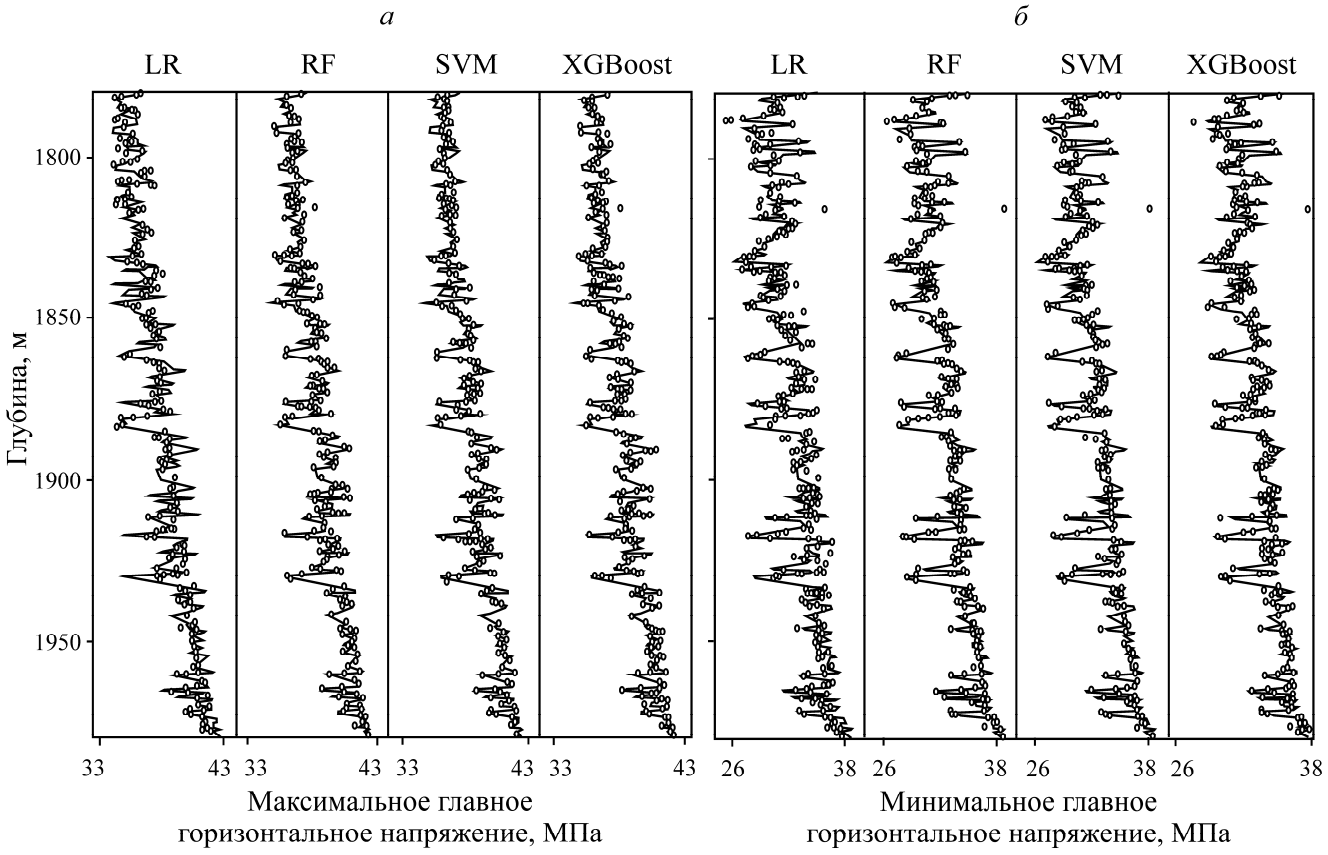

Рис. 10. Прогнозирование максимального (*а*) и минимального (*б*) главных горизонтальных напряжений рассматриваемыми моделями на основе тестового массива данных

Значения, спрогнозированные моделью XGBoost, максимально близки к фактическим и немного различаются при глубине 1800 – 1820 м. Видна значительная разница между фактическими и спрогнозированными значениями моделью линейной регрессии, особенно на глубине 1880 – 1940 м. Модель опорных векторов и модель случайного леса показали средние результаты, что соответствует результатам *k*-блочной перекрестной валидации. Точность модели XGBoost в прогнозировании максимального главного горизонтального напряжения превышает точность прогнозирования минимального напряжения, что согласуется с результатами расчета коэффициента корреляции R^2 из таблицы (для максимального главного горизонтального напряжения $R^2 = 0.975$, для минимального $R^2 = 0.944$).

На рис. 11 представлена эффективность прогнозирования поля природных напряжений в пределах заданного массива данных. Горизонтальная ось — фактические значения $y_{\text{факт}}$, вертикальная — спрогнозированные разными моделями $y_{\text{спрог}}$, сплошная линия — равенство фактических и спрогнозированных значений, т. е. чем ближе точка к данной линии, тем выше точность прогноза. Значения, спрогнозированные моделью линейной регрессии, наиболее отклоняются от фактических, что подтверждается максимальными MAE (0.493 и 0.436). Значения, спрогнозированные моделью XGBoost, отклоняются от фактических менее всего, что подтверждается минимальными значениями MAE (0.185 и 0.248). Значения моделей опорных векторов и случайного леса имеют средние результаты.

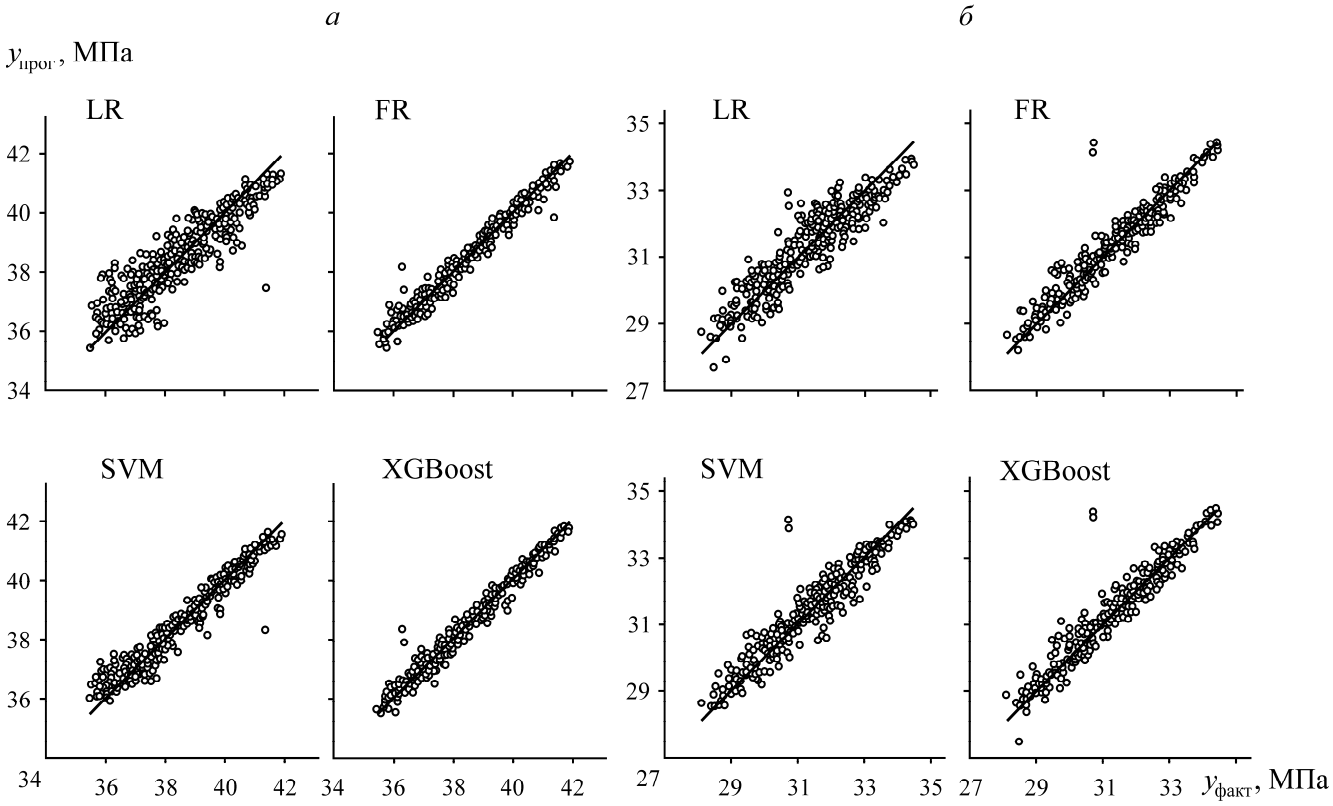

Рис. 11. Точность прогнозирования максимального (*а*) и минимального (*б*) главных горизонтальных напряжений моделями линейной регрессии (LR), случайного леса (RF), опорных векторов (SVM) и XGBoost

ВЫВОДЫ

В соответствии с методом корреляции Пирсона, характерными параметрами для определения минимального главного горизонтального напряжения являются глубина, GR, LLD, ILD, AC, VCA, для максимального — глубина, плотность, GR, SP, CAL. Модель линейной регрессии имеет наименьшее время расчета для максимального (0.00101 с) и минимального (0.00099 с) главных горизонтальных напряжений, но при этом дает наименьшую точность ($R^2 = 84.2$ и 83.8 %). Модель XGBoost сочетает высокую скорость расчета (0.16002 и 0.16895 с) и наибольшую точность ($R^2 = 97.5$ и 94.4 %). Результаты *k*-блочной перекрестной валидации выявили, что модель XGBoost обладает высокой генерализирующей способностью и надежностью. Модели опорных векторов и случайного леса дают средние результаты. В рамках тестового массива данных модель XGBoost также показала наибольшую точность. Последнее подтверждает, что рассмотренный метод исследования обладает определенной универсальностью и может быть расширен на решение других проблем, связанных с прогнозированием природных напряжений в массиве.

СПИСОК ЛИТЕРАТУРЫ

1. **Bai X., Zhang D., Wang H., Li S., and Rao Z.** A novel in situ stress measurement method based on acoustic emission Kaiser effect: A theoretical and experimental study, Royal Soc. Open Sci., 2018, Vol. 5, No. 10. — P. 488 – 504.
2. **Ma N.** Study on method and application of geostress prediction with seismic data, China University of Petroleum, East China, 2020.
3. **Ito T., Funato A., Lin W., Doan M., Boutt D. F., Kano Y., Ito H., Saffer D., McNeill L. C., Byrne T., and Moe K. T.** Determination of stress state in deep subsea formation by combination of hydraulic fracturing in situ test and core analysis: A case study in the IODP expedition 319, J. Geoph. Res.: Solid Earth, 2013, Vol. 118, No. 3. — P. 1203 – 1215.
4. **Chang C., Jo Y., Oh Y., Lee T. J., and Kim K. Y.** Hydraulic fracturing in situ stress estimations in a potential geothermal site, Seokmo Island, South Korea, J. Rock Mech. Rock Eng., 2014, Vol. 47, No. 5. — P. 1793 – 1808.
5. **Qin X., Chen Q., Wu M., Tan C., Feng C., and Meng W.** In situ stress measurements along the Beichuan – Yingxiu fault after the wenchuan earthquake, Eng. Geol., 2015, Vol. 194. — P. 114 – 122.
6. **Zhao X. G., Wang J., Cai M., Ma L. K., Zong Z. H., Wang X. Y., Su R., Chen W. M., Zhao H. G., Chen Q. C., An Q. M., Qin X. H., Ou M. Y., and Zhao J. S.** In situ stress measurements and regional stress field assessment of the Beishan area, China, Eng. Geol., 2013, Vol. 163. — P. 26 – 40.
7. **Han Z., Wang C., Wang C., Zou X., Jiao Y., and Hu S.** A proposed method for determining in situ stress from borehole breakout based on borehole stereo-pair imaging technique, Int. J. Rock Mech. Min. Sci., 2020, Vol. 127. — 104215.
8. **Thorsen K.** In situ stress estimation using borehole failures — even for inclined stress tensor, J. Petroleum Sci. Eng., 2011, Vol. 79, No. 3 – 4. — P. 86 – 100.
9. **Bai X., Zhang D., Wang H., Li S., and Rao Z.** A novel in situ stress measurement method based on acoustic emission Kaiser effect: A theoretical and experimental study, Royal Soc. Open Sci., 2018, Vol. 5, No. 10. — 181263.

10. **Cai M. and Peng H.** Advance of in situ stress measurement in China, J. Rock Mech. Geotech. Eng., 2011, Vol. 3, No. 4. — P. 373 – 384.
11. **Liu S. and Harpalani S.** Evaluation of in situ stress changes with gas depletion of coalbed methane reservoirs, J. Geoph. Res., Solid Earth, 2014, Vol. 119, No. 8. — P. 6263 – 6276.
12. **Ge X. and Hou M.** Principle of in situ 3D rock stress measurement with borehole wall stress relief method and its preliminary applications to determination of in situ rock stress orientation and magnitude in Jinping hydropower station, Sci. China Technol. Sci., 2012, Vol. 55, No. 4. — P. 939 – 949.
13. **Tao Q. and Ghassemi A.** Poro-thermoelastic borehole stress analysis for determination of the in situ stress and rock strength, Geothermics, 2010, Vol. 39, No. 3. — P. 250 – 259.
14. **Sokol Ľ., Melichar R., and Baroň I.** Present-day stress inversion from a single near-surface fault: A novel mathematical approach, J. Struct. Geol., 2018, Vol. 117. — P. 163 – 167.
15. **Ljunggren C., Chang Y., Janson T., and Christiansson R.** An overview of rock stress measurement methods, Int. J. Rock Mech. Min. Sci., 2003, Vol. 40, No. 7 – 8. — P. 975 – 989.
16. **Blanton T. L.** The relation between recovery deformation and in situ stress magnitudes, SPE, 1983.
17. **Ou Z., Li Z. J., Yang J. F., Wang T., and Zou M.** Application research on Hoek – Brown criterion in initial ground stress field evaluation, Chinese J. Underground Space Eng., 2017, Vol. 13, No. 2. — P. 387 – 401.
18. **Qiu B., Sengupta M., and Herwanger J.** Stress induced seismic velocity anisotropy study, Soc. Exp. Geoph., 2010.
19. **Ju W., Niu X. B., Feng S. B., You Y., Xu K., Wang G., and Xu H.-R.** Predicting the present-day in situ stress distribution within the Yanchang formation Chang 7 shale oil reservoir of Ordos basin, Central China, Petroleum Sci., 2020, Vol. 17, No. 4. — P. 912 – 924.
20. **Pan L., Liu H. J., Liu Y., and Li Y. J.** Application of In-stress prediction technology based on seimic method in Nanchuan district, China Sciencepaper, 2021, Vol. 16, No. 1. — P. 38 – 43.
21. **Huang J. X., Peng S. M., Wang X. J., and Xiao K.** Applications of imaginglogging data in the research of fracture and ground stress, Acta Petrolei Sinica, 2006, Vol. 27, No. 6. — P. 65 – 69.
22. **Liu J. W.** Application of multipole acoustic logging data in hard formation of Liaohe oilfield, Geoscience, 2004, Vol. 18, No. 3. — P. 378 – 382.
23. **Fu H. C., Qin W. Q., and Zhang L.** Approach to in situ stress of carbonate reservoirs in eastern Lungu of Tarim basin with well logs, Xinjiang Petroleum Geol., 2005, Vol. 26, No. 4. — P. 424 – 425.
24. **Lin Y. S., Ge H. K., and Wang S. C.** Testing study on dynamic and static elastic parameters of rocks, Chinese J. Rock Mech. Eng., 1998, Vol. 17, No. 2. — P. 106 – 112.
25. **Nian T., Wang G., Xiao C., Zhou L., Sun Y., and Song H.** Determination of in situ stress orientation and subsurface fracture analysis from image-core integration: an example from ultra-deep tight sandstone (BSJQK formation) in the Kelasu belt, Tarim basin, J. Pet. Sci. Eng., 2016, Vol. 147. — P. 495 – 503.
26. **Ju W. and Wang K.** A preliminary study of the present-day in situ stress state in the Ahe Tight Gas Reservoir, Dibei Gasfield, Kuqa Depression, Marine Pet. Geol., 2018, Vol. 96. — P. 154 – 165.
27. **Ferrero A. M., Migliazza M., Segalini A., and Gullì D.** In situ stress measurements interpretations in large underground marble quarry by 3D modeling, Int. J. Rock Mech. Min. Sci., 2013, Vol. 60. — P. 103 – 113.
28. **Nemcik J., Gale W. J., and Fabjanczyk M. W.** Methods of interpreting ground stress based on underground stress measurements and numerical modelling, Coal Operators' Conf., 2006.
29. **Wang X. F., Tang S. H., Xie H., Li Z. C., and Huang J. H.** Numerical simulation research on propagation of hydraulic fractures of coal reservoir in South Qinshui basin, Geoscience, 2012, Vol. 26, No. 3. — P. 527 – 532.

30. Zhang Y. T. and Huang H. C. Trend analysis of residual stress distribution in rock mass, Shuili Xuebao, 1984, Vol. 4. — P. 31 – 38.
31. Karatela E. and Taheri A. Localized stress field modelling around fractures using three-dimensional discrete element method, J. Pet. Sci. Eng., 2018, Vol. 171. — P. 472 – 483.
32. Feng P., Li S., Tang D. Z., Chen B., and Zhong G. H. Application of support vector machine in prediction of coal seam stress, Geoscience, 2022, Vol. 36, No. 5. — P. 1333 – 1340.
33. Shang F. H., Wang W. Q., and Cao M. J. Shale In-situ stress prediction model based on improved BP neural Network, Computer Technol. Development, 2021, Vol. 31, No. 7. — P. 164 – 170.
34. Li Y. W., Li Y. W., Shao L. F., Tian F. C., and Tang J. Z. A new physics-informed method for the fracability evaluation of shale oil reservoirs, Coal Geol. Exp., 2023, Vol. 51, No. 10. — P. 37 – 51.

Поступила в редакцию 06/II 2024
После доработки 02/III 2024
Принята к публикации 14/III 2024