https://doi.org/10.51890/2587-7399-2023-8-1-23-29

ГЕОЛОГИЯ И ГЕОЛОГО-РАЗВЕДОЧНЫЕ РАБОТЫ

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КОЛИЧЕСТВЕННОГО ПРОГНОЗА ПО ДАННЫМ СЕЙСМОРАЗВЕДКИ

А.В. Буторин
Научно-Технический Центр «Газпром нефти» (ООО «Газпромнефть НТЦ»), РФ, Санкт-Петербург

Электронный адрес: Butorin.av@gazpromneft-ntc.ru

Введение. Одной из ключевых задач для сейсморазведки является прогноз геологического строения изучаемых пластов. В частности, оценка мощности коллекторов на основании имеющейся скважинной статистики. Подобная задача является стандартной в рамках динамического анализа волнового поля и зачастую решается путем построения прогнозной модели на основании имеющейся геолого-геофизической информации, в том числе по известным значениям эффективной мощности в скважинах.

Цель. Оценка эффективности методов машинного обучения при решении задачи прогноза мощности коллекторов по данным сейсморазведки. Современный анализ данных зачастую использует эту категорию методов для построения различных прогнозных моделей. Сейсмическая интерпретация, в свою очередь, связана с использованием относительно простых линейных моделей. Это делает актуальным определение прироста качества от использования сложных моделей предсказания.

Материалы и методы. Для исследования использован относительно хорошо изученный бурением участок одного из месторождений в Западной Сибири. Рассматриваемая территория полностью покрыта данными 3D-сейсморазведки, для построения модели имеются 170 скважин, в которых определено значение эффективной мощности.

В рамках исследования рассмотрен как стандартный подход с применением линейной регрессии, так и более сложные алгоритмы машинного обучения, такие как многомерная регрессия, метод случайного леса, метод ближайших соседей и нейронная сеть. Для оценки качества предсказания имеющаяся выборка скважин разделена на обучающую и валидационную, состоящие из 80 и 90 скважин соответственно.

Все вычисления реализованы с использованием открытых библиотек языка программирования python.

Результаты. Получены распределения ожидаемой точности прогноза для каждого из рассмотренных методов. В тексте статьи подробно описан алгоритм работы, а также выполненные тесты для подбора параметров каждого алгоритма.

Заключение. Полученные результаты позволяют сделать вывод об эффективности использования методов машинного обучения. Все рассмотренные сложные алгоритмы позволяют получить более точный прогноз эффективной мощности по сравнению с подходом линейной регрессии. Наиболее значительный прирост точности наблюдается при использовании нейронной сети и составляет 23 %.

Ключевые слова: количественный прогноз, динамический анализ, машинное обучение

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Буторин А.В. Сравнение эффективности методов машинного обучения для решения задачи количественного прогноза по данным сейсморазведки. PROНЕФТЬ. Профессионально о нефти. 2023;8(1):23–29. https://doi.org/10.51890/2587-7399-2023-8-1-23-29

Статья поступила в редакцию 17.11.2022
Принята к публикации 23.12.2022
Опубликована 31.03.2023

COMPARISON OF THE EFFECTIVENESS OF MACHINE-LEARNING METHODS FOR SOLVING THE PROBLEM OF QUANTITATIVE PREDICTION BASED ON SEISMIC DATA

Aleksandr V. Butorin
Gazprom-neft STC LLC, RF, Saint Petersburg

E-mail: Butorin.av@gazpromneft-ntc.ru

Introduction. One of the key tasks for seismic interpretation is the prediction of the geological structure of the studied formations. In particular, a common task is to estimate the net thickness of reservoirs based on available well statistics. Such a task is standard in the framework of dynamic wave field analysis and is often solved by constructing a predictive model based on available geological and geophysical information, including values of net thickness in available wells.

Goal. The purpose of the work is to evaluate the effectiveness of machine learning methods in solving the problem of reservoir thickness prediction based on seismic data. Modern data analysis often uses this category of methods to build various predictive models. Seismic interpretation, in turn, is often associated with the use of relatively simple linear models. This makes it relevant to determine the gain from the use of complex prediction models.

Materials and methods. To carry out the study, a relatively well-studied area of one of the fields in Western Siberia was used. The territory under consideration is completely covered with 3D seismic data, there are 170 wells for constructing the model, in which the value of net thickness is determined.

To implement the study, both a standard linear regression and more complex machine learning algorithms are considered. Among the algorithms, multidimensional regression, random forest method, nearest neighbor method and neural network are considered. To assess the quality of prediction, the available sample of wells is divided into training and validation samples consisting of 80 and 90 wells, respectively.

All calculations are implemented using open python programming language libraries.

Results. As a result, distributions of the expected accuracy of the forecast for each of the considered methods were obtained. The text of the article describes in detail the research algorithm, as well as the tests performed to select the parameters of each algorithm.

Conclusion. The results obtained allow us to conclude about the effectiveness of using machine-learning methods. All the approaches considered make it possible to obtain a more accurate prediction of the net thickness compared to the linear regression approach. The most significant increase in accuracy is observed with using a neural network and the improvement estimated as 23 %.

Key words: quantitative analysis, dynamic analysis, machine learning

Conflict of interest: the authors declare no conflict of interest.

For citation: Butorin A.V. Comparison of the effectiveness of machine-learning methods for solving the problem of quantitative prediction based on seismic data. PRONEFT. Professionally about oil. 2023;8(1):23–29. https://doi.org/10.51890/2587-7399-2023-8-1-23-29

Manuscript received 17.11.2022
Accepted 23.12.2022
Published 31.03.2023

ВВЕДЕНИЕ

Ключевой задачей интерпретации данных сейсморазведки является прогноз геологических параметров среды по характеристикам волнового поля. К подобным параметрам могут быть отнесены: глубина залегания пласта, его литологический состав, песчанистость, пористость проницаемость, насыщенность. Характеристики волнового поля могут быть разделены на кинематические, связанные с временем регистрации отражения, и динамические, связанные с его энергетическими параметрами. В рамках данного исследования рассмотрена задача прогноза мощности коллектора по динамическим характеристикам волнового поля.

Выбранная задача является стандартной и выполняется в большинстве геолого-геофизических проектов при наличии необходимой скважинной информации, что обуславливает актуальность данного исследования. Связь между параметрами амплитуды и эффективной мощностью коллектора — хорошо изученный феномен, основанный на интерференции отражений. Одна из известных публикаций [1] наглядно показывает на модельных данных сублинейную связь между мощностью песчаника с амплитудой волнового поля. В целом для большинства практических проектов подход остается схожим и заключается в поисках наиболее достоверной линейной связи между поисковым параметром (эффективной толщиной) и характеристикой амплитуды (атрибутом). Найденная функциональная связь в дальнейшем используется для прогноза искомого параметра в межскважинном пространстве.

С точки зрения алгоритмов машинного обучения рассматриваемая задача относится к категории алгоритмов «обучения с учителем» — априорно известные значения в точках скважин и набор атрибутов волнового поля формируют обучающую выборку, которая используется для создания предсказывающей модели. С этой точки зрения рассматриваемая геологическая задача является стандартной задачей регрессии в рамках машинного обучения, решаемая с помощью множества алгоритмов, реализованных в различных языках программирования. В рамках данного исследования все вычисления выполнены с использованием открытых библиотек языка программирования python, в частности библиотеки sklearn, содержащей реализации основных алгоритмов машинного обучения.

В качестве объекта исследования выбран разбуренный участок одного из месторождений в Ханты-Мансийском автономном округе Западной Сибири. Целевыми пластами в рамках рассматриваемого района являются пласты группы АС, сформированные в условиях мелководно-морских обстановок. Развитие коллектора связано с фациями аккреционной системы меандрирующего русла, однозначно картируемого по данным сейсморазведки. Всего на участке пробурено 170 скважин, вся территория участка покрыта данными сейсморазведки МОГТ 3D с кратностью системы наблюдения 144 (**рис. 1**).

МЕТОДЫ

На начальном этапе имеющиеся скважины были разделены на две выборки: 80 скважин использованы в качестве обучающего массива, 90 скважин исключены из процесса построения предсказывающей модели и использовались на финальном этапе в качестве отложенной выборки для оценки качества каждой модели. Подобный подход является необходимым условием при работе с алгоритмами машинного обучения, наличие валидационной выборки позволяет объективно оценить качество и проконтролировать отсутствие эффекта переобучения, то есть его настройки на имеющуюся обучающую выборку.
В рамках исследования рассмотрены наиболее распространенные методы, доступные для применения в библиотеке sklearn: многомерная регрессия, метод случайного леса, метод ближайших соседей и нейронная сеть. Все эти методы являются хорошо известными и подробно описаны в научной литературе, по этой причине в данной работе не приводятся их математические формулировки и не описываются алгоритмы вычисления.
Для формирования обучающей выборки использованы трассы исходного суммарного куба в точках скважин. Вычисление атрибутов для обучающей выборки выполнено в интервале 40 мс относительно целевого отражающего горизонта, соответствующего кровле продуктивного пласта. В рамках заданного окна были вычислены стандартные атрибуты волнового поля: мгновенная амплитуда, а также магнитуды по частотным компонентам 10, 20, 30 и 40 Гц, полученные с использованием непрерывного вейвлет-преобразования по вейвлетам Риккера. Внутри рассматриваемого интервала для оценки характеристик атрибутов выполнен перебор окон от 0 до 8 мс. Внутри каждого окна использованы различные статистические оценки: сумма, минимальное и максимальное значение, среднеквадратическое значение. Данные вычисления выполнены для каждой трассы, что позволило получить на выходе массив из 1800 атрибутов для каждой скважины. Набор атрибутов и истинные значения эффективной мощности формируют входной массив для обучения алгоритмов.
Полученный массив значений, вычисленных по трассам суммарного куба, характеризуется высокой степенью корреляции для некоторых атрибутов. Данное обстоятельство может негативно сказываться на ходе обучения некоторых алгоритмов, поэтому для минимизации этого фактора к атрибутам

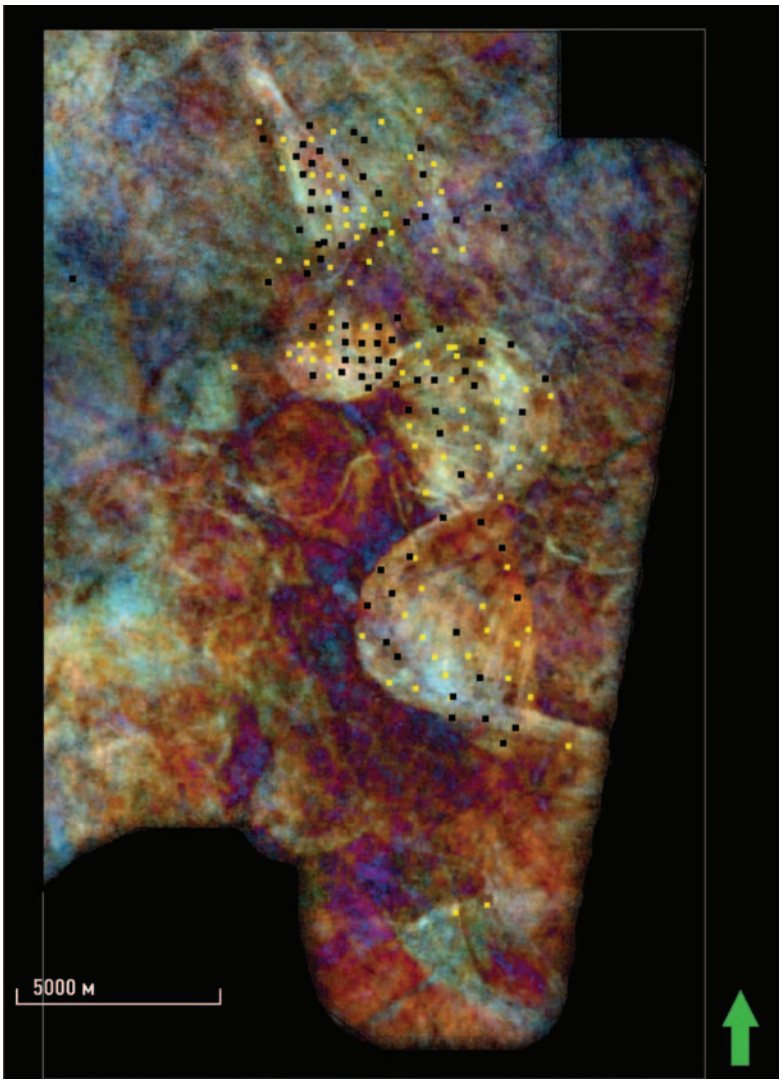

Рис. 1. Карта RGB-смешивания по целевому пласту. Черные точки — выборка скважин для обучения; желтые точки — скважины для валидации (А.В. Буторин)
Fig. 1. RGB-map for aiming horizon. Black dots — sample wells for education of model; yellow dots — sample for validation (Aleksandr V. Butorin)

применен метод главных компонент. Метод главных компонент (PCA) позволяет получить некоррелируемые атрибуты, однако количество компонент определяется эмпирически и в ходе исследования данное значение варьировалось от 1 до 50. По результатам тестирования отмечено, что применение метода главных компонент позволяет, с одной стороны, повысить скорость обучения, а с другой стороны — положительно влияет на метрики качества. Дополнительно протестировано добавление к обучающей выборке результатов кластеризации по форме трассы. Кластеризация выполнена с применением метода К-средних, представленного в библиотеке sklearn. Установлено, что добавление результатов кластеризации не влияет на качество моделей.
Таким образом, был сформирован массив для дальнейшего применения методов

количественного прогноза. Массив представлен истинными значениями эффективной мощности по 80 скважинам, а также набором атрибутов после применения метода главных компонент с выбором от 1 до 50 элементов для каждой точки скважины. Данный массив использовался для обучения моделей машинного обучения и получения прогнозного алгоритма, который в дальнейшем применялся к валидационной выборке с оценкой среднеквадратического отклонения (СКО) прогноза.

РЕЗУЛЬТАТЫ

На начальном этапе получены оценки базовых алгоритмов прогноза. В качестве одного из алгоритмов протестирован метод среднего значения. В рамках данной модели для каждой скважины валидационной выборки прогнозное значение принималось равным среднему по обучающей выборке. Подобный алгоритм позволяет получить среднеквадратическую ошибку, равную 5,91 м. Данная оценка является наиболее грубой моделью, при которой не учитываются имеющиеся сейсмические данные, и она может использоваться как фундамент для последующей оценки эффективности более сложных алгоритмов.

В качестве второго базового алгоритма рассмотрена линейная регрессия по одному атрибуту. Для реализации данного подхода по каждому из 1800 атрибутов с использованием метода наименьших квадратов получена зависимость эффективной мощности от атрибута, которая в дальнейшем применялась к атрибутам валидационной выборки. Полученная статистика показана на **рис. 2**. Выбор условно наилучшей модели сделан по коэффициенту корреляции, который составил 0,68. При подобном подходе точность прогноза эффективной мощности на валидационной выборке составила 5,69 м. В данном случае можно оценить эффект от учета геофизических данных. В рассматриваемом примере прирост точности оказывается незначительным, что связано с достаточно однозначным выделением геологического объекта и относительно высокой успешностью бурения.

Необходимо отметить, что полученный результат отражает неопределенность решения задачи количественного прогноза. Как видно из **рис. 2**, в имеющейся выборке присутствуют атрибуты, обеспечивающие более высокую точность на валидационной выборке, — минимальная ошибка составляет 4,96 м. Однако данный атрибут характеризуется меньшим значением коэффициента корреляции на обучающей выборке, поэтому его выбор невозможен. Данный факт хорошо иллюстрирует неопределенности, обусловленные ограниченностью имеющейся скважинной статистики.

Таким образом, получены начальные оценки точности прогноза эффективной мощности при использовании относительно простых алгоритмов. Данные значения применяются

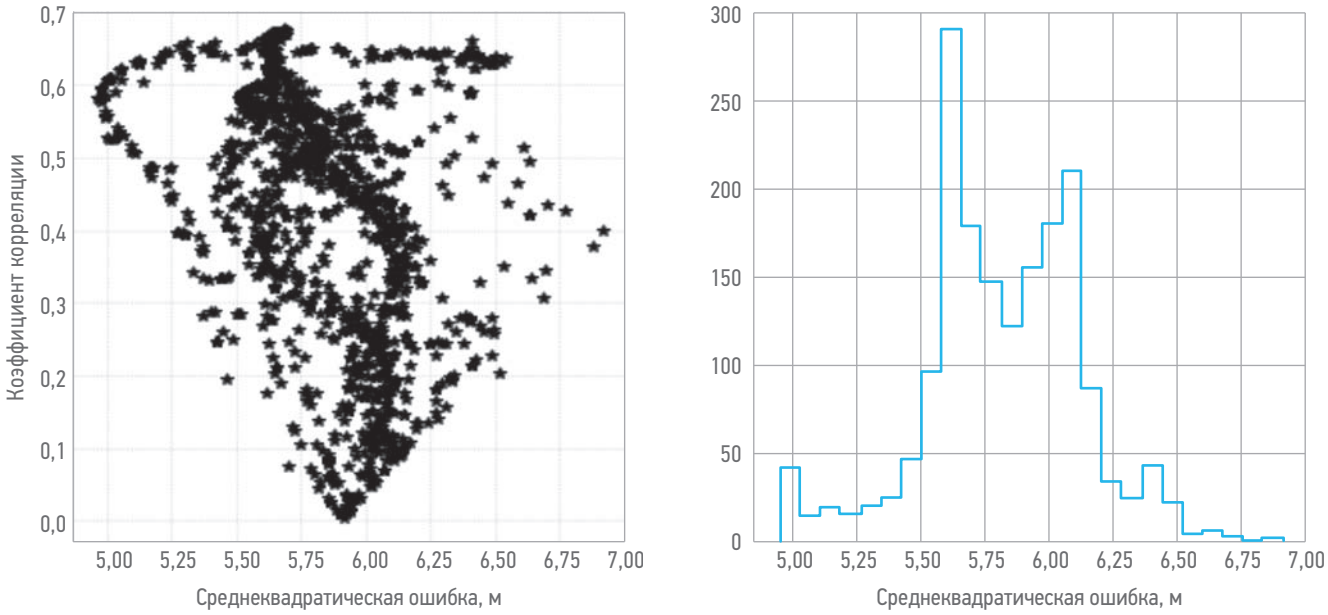

Рис. 2. Статистика применения линейной регрессии по одному атрибуту. Слева — кросс-плот между коэффициентом корреляции и среднеквадратической ошибкой; справа — гистограмма среднеквадратических ошибок (А.В. Буторин)

Fig. 2. Statistics of linear regression for single attribute. Left — cross-plot for correlation coefficient and standard deviation of errors. Right — histogram of standard deviation of errors (Aleksandr V. Butorin)

в дальнейшем для оценки эффективности методов машинного обучения.
Рассмотрим алгоритм исследования для методов машинного обучения: на начальном этапе используемая выборка случайным образом разделялась на обучающую и тестовую в пропорции 70 и 30 %. Первая выборка использовалась для обучения алгоритма, вторая — для оценки метрик качества и выбора наилучшей модели. В качестве критерия выбора использовался коэффициент корреляции фактического и прогнозного значения эффективной мощности. В ходе обучения создавался цикл вычисления главных компонент от 1 до 50 с шагом 5. Внутри цикла происходило обучение и выбор наилучшей модели для выбранного алгоритма. Наиболее точный алгоритм применялся к отложенной валидационной выборке для оценки среднеквадратического отклонения прогноза. Описанная последовательность действий повторялась 100 раз для получения распределения ошибки прогноза для каждого из методов. За счет использования случайного разделения выборки и вероятностной природы алгоритмов каждая реализация характеризовалась своим итоговым значением среднеквадратической ошибки прогноза, что позволило сформировать статистику по эффективности каждого алгоритма.
Таким образом, для каждого из рассматриваемых методов машинного обучения была сформирована гистограмма, показывающая ожидаемый диапазон точности прогноза.

ОБСУЖДЕНИЕ

Рассмотрим результаты применения методов машинного обучения. На **рис. 3** приведены гистограммы для каждого из используемых алгоритмов, включая линейную регрессию с одним параметром. Необходимо отметить, что в данном случае не производился отбор моделей простой регрессии по коэффициенту корреляции, поэтому полученное распределение имеет большую дисперсию и бимодальную форму — левый максимум соответствует моделям с высокой корреляцией, правый — моделям с низкой корреляцией.
Наиболее простым алгоритмом является многомерная регрессия [2]. В общей постановке использована регрессия с регуляризацией. Регрессия использует линейную комбинацию входных атрибутов для построения прогнозной модели. При этом регуляризация позволяет дополнительно ввести штраф на значение мультипликатора при каждом атрибуте, чтобы избежать слишком больших значений весового коэффициента. Без использования регуляризатора регрессия представляет собой применение метода

ИСПОЛЬЗОВАНИЕ НЕЙРОННОЙ СЕТИ ДЛЯ ПРОГНОЗА МОЩНОСТИ КОЛЛЕКТОРОВ ПО ДАННЫМ СЕЙСМОРАЗВЕДКИ ПОВЫШАЕТ ТОЧНОСТЬ НА 23 % ПО СРАВНЕНИЮ С ДРУГИМИ МЕТОДАМИ.

наименьших квадратов. В рамках исследования рассмотрено два вида регуляризации: минимизация суммы квадратов весовых коэффициентов — Ridge и минимизация суммы абсолютных значений — Lasso.
Принципиальным отличием этих алгоритмов является возможность обнуления некоторых атрибутов в рамках Lasso-регрессии.
Для регрессии по методу наименьших квадратов точность модели варьируется от 5,02 до 6,20 м с математическим ожиданием 5,49 м. Ridge-регрессия показывает аналогичное распределение точности. Lasso-регрессия показывает бóльшую точность: минимальное значение составило 4,75 м, максимальное — 5,39 м, математическое ожидание — 5,12 м.
Таким образом, использование Lasso-регрессии позволяет повысить прогнозную точность по отношению к использованию регрессии по одному атрибуту.
Другим алгоритмом машинного обучения выступал метод случайного леса [3], заключающийся в использовании ансамбля решающих деревьев для формирования предсказывающей модели. Ключевыми настройками

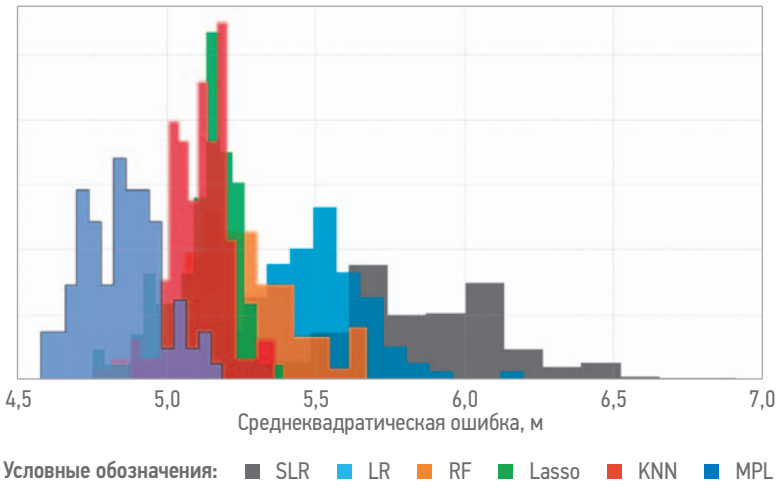

Рис. 3. Полученные распределения ошибок для рассматриваемых методов (SLR — регрессия по одному атрибуту, LR — многомерная регрессия, RF — случайный лес, Lasso-регрессия, KNN — метод ближайших соседей, MPL — нейронная сеть) (А.В. Буторин)
Fig. 3. Histograms of standard deviation of errors for different methods (SLR — single linear regression, LR — linear regression, RF — random forest, Lasso-regression, KNN — method of nearest neighbors, MPL — neural net) (Aleksandr V. Butorin)

данного алгоритма является количество решающих деревьев и их глубина. Для подбора указанных гиперпараметров использован метод перебора по сетке — глубина варьировалась от 2 до 10 разбиений, а количество деревьев — от 5 до 40. Полученное распределение ошибки характеризуется минимальным значением 4,76 м, максимальным — 5,67 м, с математическим ожиданием — 5,24 м.
По результатам расчетов метод случайного леса показал меньшую точность, чем Lasso-регрессия.
Следующим рассмотренным алгоритмом является метод ближайших соседей [4], который использует осреднение по заданному количеству соседних точек обучения. Для определения соседних точек используется евклидово расстояние в области атрибутов. Ключевым значимым параметром модели выступает количество соседних точек для осреднения, в данном случае также выполнен перебор значений от 2 до 20 точек. Как показало тестирование, оптимальное количество соседей находится в диапазоне 10–16 точек. Полученное распределение по 100 реализациям позволяет оценить распределение точности: минимальное значение ошибки — 4,88 м, максимальное — 5,37 м, математическое ожидание ошибки — 5,12 м. Как видно из гистограммы, ожидаемая точность прогноза методом ближайших соседей практически совпадает со значением точности Lasso-регрессии.
Последним из рассматриваемых алгоритмов выступала двухслойная нейронная сеть [5]. В рамках нейронной сети создается последовательность слоев, на каждом из которых осуществляется линейная комбинация входных значений атрибутов с применением заданной функции активации. Дополнительно для стабилизации решения использована регуляризация по сумме квадратов весовых коэффициентов. Как показало тестирование, наилучший результат достигается

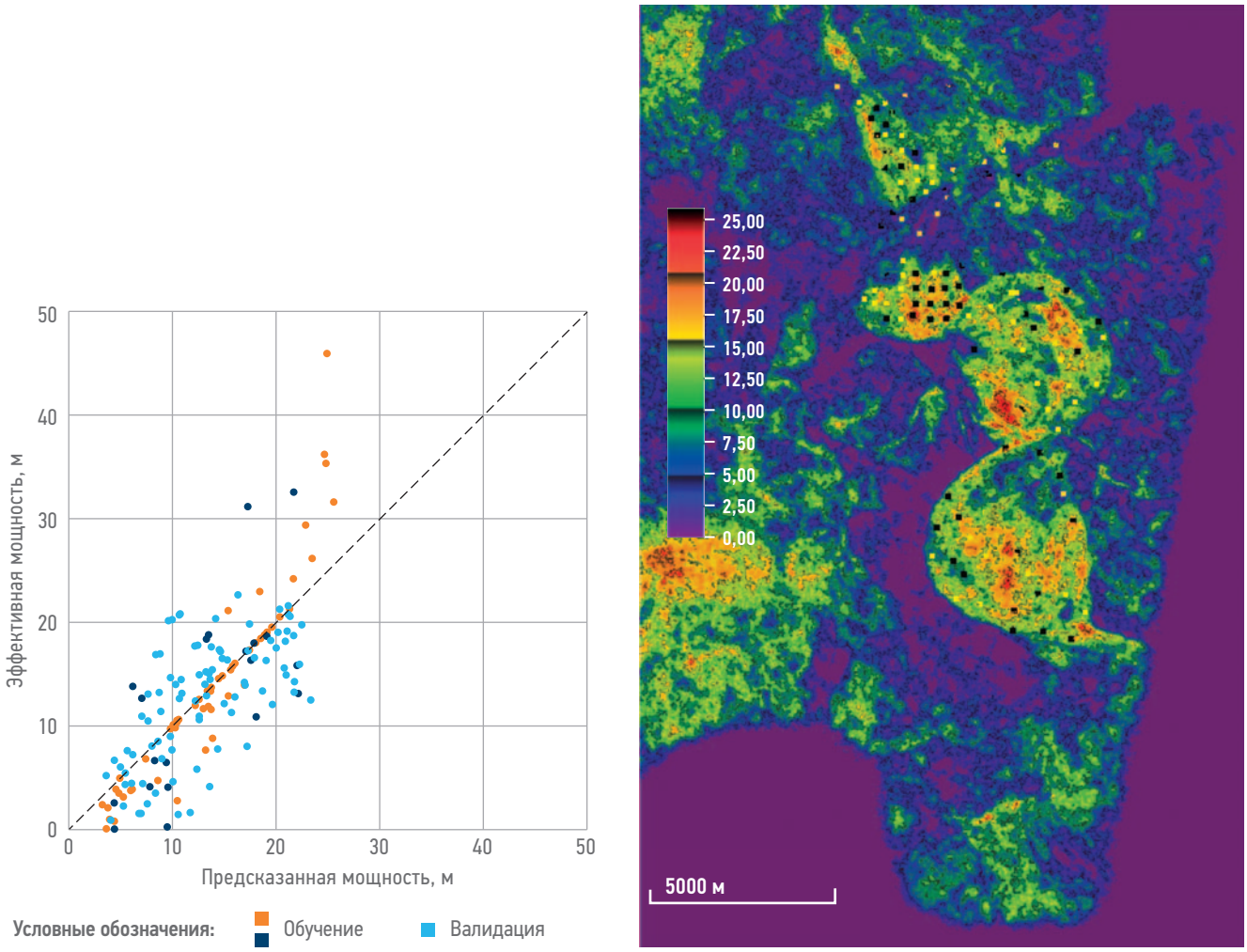

Рис. 4. Результат прогноза эффективной мощности с применением алгоритма нейронной сети. Слева — сопоставление фактических и прогнозных значений (пунктиром показана линия y=x, красные точки — обучающая выборка, синие точки — тестовая выборка на этапе обучения, зеленые точки — валидационная выборка); справа — прогнозная карта эффективной мощности пласта (А.В. Буторин)
Fig. 4. Result of reservoir thickness estimation from neural network. Left — comparison of true and predicted values (dotted line — y=x, red dots — train set, blue dots — test set, green dots — validation set); Right — map of reservoir thickness (Aleksandr V. Butorin)

при использовании логистической функции активации. Полученное распределение точности характеризуется следующими параметрами: минимальное значение — 4,58 м, максимальное — 5,19 м, математическое ожидание ошибки — 4,85 м.

Таким образом, использование нейронной сети позволяет получить наименьшую прогнозную оценку ошибки. Прирост в качестве прогноза составляет 23 % по отношению к методу регрессии по одному атрибуту. Анализируя результат применения одной из моделей, основанной на нейронной сети (**рис. 4**), можно отметить, что дисперсия на обучении и валидации остается схожей, что говорит об отсутствии переобучения модели. Однако отмечается занижение прогноза в областях с высокой мощностью, что может быть связано с недостаточностью подобных скважин в тестовой выборке.

Резюмируя проведенное исследование, можно отметить, что использование методов машинного обучения позволяет повысить точность прогноза на 23 %. Наилучший результат достигается при использовании нейронной сети, в то время как остальные методы показывают схожую точность прогноза и обеспечивают повышение точности прогноза около 10–13 %. Незначительный прирост в информативности связан с некоторой смещенностью имеющейся статистики — бурение скважин априорно ориентировалось на выделяемый геологический объект, что приводит к удовлетворительному прогнозу даже при использовании модели среднего значения. Результаты исследования показывают простоту использования современных алгоритмов анализа, все вычисления выполнены с применением открытых библиотек языка python.

Список литературы / References

1. Meckel L.D., Nath A.K. Geologic considerations for stratigraphic modelling and interpretation. In Seismic Stratigraphy — Applications to Hydrocarbon Exploration, ed. C. E. Payton. AAPG Memoir, 1977, no. 26, pp. 417–438.
2. Linear Models [Electronic Resource]. Access: https://scikit-learn.org/stable/modules/linear_model.html#lasso
3. Decision Trees [Electronic Resource]. Access: https://scikit-learn.org/stable/modules/tree.html#regression
4. Nearest Neighbors [Electronic Resource]. Access: https://scikit-learn.org/stable/modules/neighbors.html
5. Neural network models (supervised) [Electronic Resource]. Access: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

ВКЛАД АВТОРА / AUTHOR CONTRIBUTIONS

А.В. Буторин — разработал концепцию исследования, подготовил текст и рисунки. Согласен принять на себя ответственность за все аспекты работы.

Aleksandr V. Butorin — developed the article concept, prepared the text and pictures. Accepted the responsibility for all aspects of the work.

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Александр Васильевич Буторин — кандидат геолого-минералогических наук, доцент кафедры «Геофизика» Института наук о Земле СПбГУ, руководитель по развитию дисциплины «сейсморазведка» ООО «Газпромнефть НТЦ»
190000, Россия, г. Санкт-Петербург, Набережная реки Мойки, д. 75–79, литер Д
e-mail: Butorin.AV@gazpromneft-ntc.ru
AuthorID: 877389
SPIN-код: 8474-6120
Web of Science: B-7405-2019
ORCID: https://orcid.org/0000-0002-6074-1439
Scopus: 56370048400

Aleksandr V. Butorin — Cand. Sci. (Geol.-Min.), Associate Professor at the Department of Geophysics at Institute of Earth Sciences, Head of seismic discipline Gazpromneft STC LLC
75–79 liter D, Moika River emb., 190000, Saint Petersburg, Russia.
e-mail: Butorin.AV@gazpromneft-ntc.ru
AuthorID: 877389
SPIN-code: 8474-6120
Web of Science: B-7405-2019
ORCID: https://orcid.org/0000-0002-6074-1439
Scopus: 56370048400