УДК 622.276.66
DOI: 10.18799/24131830/2024/5/4428
Шифр специальности ВАК: 01.02.01

Отбор скважин-кандидатов при обработке призабойной зоны пласта методами машинного обучения

М.А. Ямкин[1]✉, Е.У. Сафиуллина[1], А.В. Ямкин[2]

[1] Санкт-Петербургский горный университет, Россия, г. Санкт-Петербург
[2] «ООО Газпром трансгаз Томск», Россия, г. Томск

✉makson.yamkin@mail.ru

Аннотация. ***Актуальность*** данной работы связана с тем, что в настоящее время на месторождениях нефти широко применяются различные технологии по увеличению нефтеотдачи и интенсификации притока, такие как обработка призабойной зоны соляной кислотой. В связи с массовым применением данной технологии на передний план выходят проблемные вопросы, в том числе связанные с выбором правильных на данный момент времени скважин-кандидатов для проведения обработки призабойной зоны. ***Цель*** данной работы заключается в оптимизации поиска скважин-кандидатов для проведения обработки призабойной зоны. В работе исследуется возможность использования моделей машинного обучения для предсказания ответа, будет ли скважина являться правильным кандидатом для проведения обработки призабойной зоны. ***Объектом*** исследования являются модели машинного обучения библиотеки sklearn. ***Методы.*** Для решения задачи предсказания, является ли скважина кандидатом для проведения обработки призабойной зоны, использовалось три модели машинного обучения библиотеки sklearn: RandomForestClassifier (далее модель обучающего леса), DecisionTreeClassifier (далее модель обучающего дерева), LinearRegression (далее модель линейной регрессии). Для оценки качества построенных моделей использовались следующие метрики той же библиотеки: F1-score, AUC-ROC-score. ***Результаты.*** Наилучший результат при обучении показала модель обучающего леса. На метрике F1-score данная модель, примененная на тестовой выборке, показала сходимость 99,5 %, а на метрике AUC-ROC-score точность составила 99,9 %. Полученная точность указывает на корректность использования модели обучающего леса для решения задачи определения правильных скважин-кандидатов. ***Заключение.*** Получена модель машинного обучения, дающая предсказание, будет ли скважина являться правильным кандидатом для проведения обработки призабойной зоны, с точностью 99,5 %.

Ключевые слова: обработка призабойной зоны пласта, скважины-кандидаты, машинное обучение, модель обучающего леса, sklearn, F1-score

Для цитирования: Ямкин М.А., Сафиуллина Е.У., Ямкин А.В. Отбор скважин-кандидатов при обработке призабойной зоны пласта методами машинного обучения // Известия Томского политехнического университета. Инжиниринг георесурсов. – 2024. – Т. 335. – № 5. – С. 7–16. DOI: 10.18799/24131830/2024/5/4428

UDC 622.276.66
DOI: 10.18799/24131830/2024/5/4428

Machine learning methods for selecting candidate wells for bottomhole formation zone treatment

M.A. Yamkin[1]✉, E.U. Safiullina[1], A.V. Yamkin[2]

[1] St. Petersburg Mining University, St. Petersburg, Russian Federation
[2] «Gazprom transgaz Tomsk LLC», Tomsk, Russian Federation

✉makson.yamkin@mail.ru

Abstract. ***Relevance.*** The fact that currently various technologies are widely used in oil fields to increase oil recovery and intensify the inflow, such as treatment of a bottomhole zone with hydrochloric acid. In relation to the widespread use of this

technology, problematic issues are coming to the fore, including those related to the selection of the right candidate wells at a given time for carrying out well treatment. ***Aim.*** To optimize the search for candidate wells for carrying out treatment of the bottomhole zone. The work explores the possibility of using machine learning models to predict whether a well will be the right candidate for a well treatment. ***Object.*** Machine learning models of the sklearn library. ***Methods.*** To solve the problem of predicting whether a well is a candidate for bottomhole treatment, three machine learning models of the sklearn library were used: RandomForestClassifier, DecisionTreeClassifier, LinearRegression. To assess the quality of the constructed models, the following metrics from the same library were used: F1-score, AUC-ROC-score. ***Results.*** The learning forest model showed the best results during training. Using the F1-score metric, this model showed 99.5% convergence on the testing dataset, and using the AUC-ROC-score metric, the accuracy was 99.9%. The resulting accuracy indicates the correctness of using RandomForestClassifier model to solve the problem of identifying the correct candidate wells. ***Conclusion.*** The machine learning model was obtained that predicts with 99.5% accuracy whether a well will be the right candidate for a well treatment.

Keywords: treatment of a bottomhole formation zone, candidate wells, machine learning, RandomForestClassifier, sklearn, F1-score

For citation: Yamkin M.A., Safiullina E.U., Yamkin A.V. Machine learning methods for selecting candidate wells for bottomhole formation zone treatment. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering*, 2024, vol. 335, no. 5, pp. 7–16. DOI: 10.18799/24131830/2024/5/4428

Введение

Технология обработки призабойной зоны (ОПЗ) в настоящее время является популярной технологией увеличения нефтеотдачи и интенсификации притока [1]. Суть данной технологии заключается в нагнетании под высоким давлением жидкостной смеси, состоящей из кислоты, воды, интенсификаторов, ингибиторов, растворителей, стабилизаторов и других компонентов [2]. Кислота, находящаяся в закачиваемой смеси, растворяет горную породу, тем самым увеличивая ее пористость и проницаемость [3, 4]. В пласт закачивают различные составы в зависимости от состава горной породы. Так, например, если в горной породе будут преобладать карбонаты, закачиваться будет соляная кислота [5].

Технология обработка призабойной зоны пласта применяется на ряде стадий разработки месторождений. Так, например, на первой и второй стадиях данная технология используется для увеличения притока путем очистки призабойной зоны пласта (ПЗП) от загрязнений, которые вызваны бурением. На третьей и четвертой стадиях разработки месторождений данная технология используется для увеличения нефтеотдачи пластов путем соединения между собой нефтенасыщенных пластов, которые разъединены различными глинистыми пропластками для улучшения их гидродинамической связи [6].

В связи с широким применением данной технологии в нефтегазовых компаниях анализируется совокупность различных параметров для выбора правильных скважин-кандидатов для проведения ОПЗ. Одним из эффективных подходов для проведения данного анализа является использование моделей машинного обучения.

Актуальность

Оптимизация процесса отбора скважин-кандидатов с целью минимизации влияния человеческого фактора, сокращения времени и исключения ошибок является актуальной проблемой. В данной работе исследуется решение данной проблемы при помощи машинного обучения. Рассматриваемая задача является задачей классификации: если скважина является правильным кандидатом для проведения ОПЗ, модель выдает ответ «1», если скважина не является правильным кандидатом для проведения ОПЗ, модель выдает «0» [7]. Объектом исследования являются скважины месторождения, расположенного в Западной Сибири, добывающие нефть с трех пластов. Данное месторождение является молодым и находится на второй стадии разработки. В связи с этим операция ОПЗ проводилась на скважине в том случае, если она не вышла на проектный режим после ее запуска в эксплуатацию. Машинное обучение применяется для скважин до попытки их вывода на режим для определения тех, которые потенциально не выйдут на режим. Это актуально в силу того, что увеличивает полезное время работы скважины, так как при проведении ОПЗ на правильных скважинах-кандидатах они сразу выйдут на режим. В качестве метода интенсификации притока или в качестве метода увеличения нефтеотдачи ОПЗ на данном месторождении не рассматривается. Скважина считается вышедшей на режим, если ее дебит соответствует рабочей характеристике насоса, динамический уровень установился на постоянной отметке и объем жидкости, отобранный из скважины, равен двум объемам ее обсадной колонны, но не менее двух объемов использованной при ремонте жидкости глушения [8].

Для создания корректной модели машинного обучения в данной работе с целью обучения выбраны признаки, которые анализируются инженерами-нефтяниками для принятия решения о назначении скважины кандидатом, чтобы прове-

сти ОПЗ. В соответствии с работой [9] для принятия решения об объявлении скважины кандидатом для проведения ОПЗ анализируются следующие факторы: тип, литология и расчлененность коллектора, степень обводненности скважины, температура пласта, дебит жидкости, скин-фактор, пластовое давление, толщина продуктивного горизонта, проницаемость.

На данный момент машинное обучение в основном применяется для решения следующих задач в нефтегазовой отрасли:

1) Модель временных рядов для предсказания пластового давления во времени. Данная модель используется вместо проведения гидродинамических исследований скважин (ГДИС) для построения кривой восстановления давления (КВД) [10, 11]. Точность данной модели составляет около 90 % [10, 11].
2) Классификационная модель для принятия решения, будет ли скважина являться кандидатом для проведения гидроразрыва пласта. Точность данной модели составила также около 90 % [12].
3) Регрессионная модель для предсказания дебита после проведения ОПЗ на скважине.

Таким образом, в литературе описано применение машинного обучения для ОПЗ только для решения регрессионной задачи. При этом использовалось простое дерево решений, описанное автором вручную [13, 14]. Для задачи, рассматриваемой в данной работе, решение с помощью методов машинного обучения не было найдено. В связи с этим в данной работе исследуется возможность использования моделей машинного обучения для предсказания ответа, будет ли скважина являться правильным кандидатом для проведения ОПЗ.

Методы исследования

В данной работе анализируются следующие модели машинного обучения: обучающего леса, обучающего дерева, линейной регрессии. Данные модели библиотеки sklearn [15] решают задачу классификации применительно к выбору скважин-кандидатов для проведения ОПЗ. Для оценки результатов обучения были использованы метрики F1-score и AUC-ROC из библиотеки sklearn [16].

На первоначальном этапе были выбраны данные для определения того, будет ли скважина являться правильным кандидатом. В соответствии с работой [14] рассматривались такие данные, как текущее пластовое давление, текущая обводненность скважины, дебит нефти, дебит жидкости, проницаемость, вязкость нефти, вязкость воды, вязкость жидкости, радиус контура питания скважины, расстояние между скважинами, диаметр эксплуатационной колонны, карбонатность коллектора, доломитизация коллектора, глинистость коллектора, инклинометрия, расчлененность, изначальное пластовое давление, нефтенасыщенность, водонасыщенность, назначение скважины, плотность воды, плотность нефти, коэффициент продуктивности, пористость, нефтенасыщенная толщина, общая толщина коллектора, плотность добываемой смеси.

Выбранные данные выгружались из отчетов по исследуемому месторождению и включали в себя файлы MS Excel, а также различные документы в формате MS Word и PDF. Глинистость, карбонатность, доломитизация, начальное пластовое давление, пористость, нефтенасыщенная и общая толщины коллектора, водо- и нефтенасыщенность на момент начала разработки, расчлененность являлись постоянными для каждого из трех пластов месторождения. В связи с этим для увеличения скорости обучения и исключения мультиколлинеарности признаков параметры, постоянные для каждого из трех пластов, определялись одним, созданным искусственно, категориальным признаком – пласт, применение которого возможно лишь для рассматриваемого месторождения. В итоге модель машинного обучения обучалась на 18 признаках. В табл. 1 представлены признаки и их названия, которые использовались для обучения. Также был выделен целевой признак – является ли скважина правильным кандидатом для проведения ОПЗ или нет. Эти данные также были взяты из файлов формата MS Excel по исследуемому месторождению. Выборка данных была разделена на тренировочную и тестовую. При анализе моделей машинного обучения использовалась кросс-валидация на тренировочной выборке с подбором гиперпараметров [14]. Кросс-валидация позволяет более точно оценить сходимость моделей машинного обучения. Это осуществляется путем разделения тренировочной выборки на несколько частей (в соотношении, которое выбирает автор) и запуска модели машинного обучения несколько раз на различных подвыборках из тренировочной выборки, и оценки ее сходимости. В данной работе тренировочная выборка была разбита на 4 подвыборки, следовательно, 4 раза вычислялось значение метрики F1-score и по этим 4 значениям вычислялось среднее, которое и приводится в дальнейшем.

Следует отметить, что при обработке целевого признака был выявлен сильный дисбаланс классов, составляющий 0,05, то есть скважин-некандидатов больше, чем скважин-кандидатов, в 20 раз. Для борьбы с этой проблемой применялись методы, описанные далее в данной статье. Кроме этого, для борьбы с дисбалансом классов делалось следующее допущение: если скважина признавалась кандидатом для проведения ОПЗ, она являлась кандидатом в течение всего последующего месяца.

Таблица 1. *Признаки для обучения*

Table 1. *Features for learning*

Признак/Feature	Название в модели/Name in model
Текущее пластовое давление Current reservoir pressure	ТР Рпласт [атм]/TR Pplast [atm]
Текущая обводненность Current watercut	Обводненность V OIS протяжка [%] Watercut V OIS broach [%]
Дебит нефти Oil flow rate	Qн OIS протяжка [тн/сут] Qo OIS broach [t/d]
Дебит жидкости Liquid flow rate	Qж OIS протяжка [м³/сут] Ql OIS broach [m³/d]
Проницаемость Permeability	ТР Проницаемость [мД] TR Permeability [mD]
Вязкость нефти Oil viscosity	ТР Вязкость нефти в пл. усл. [спз] TR Oil viscosity in reservoir condition [spz]
Вязкость воды Water viscosity	ТР Вязкость воды в пл. усл. [спз] TR Water viscosity in reservoir condition [spz]
Вязкость жидкости Liquid viscosity	ТР Вязкость жидкости [спз] TR Liquid viscosity [spz]
Радиус контура питания скважины Well feed circuit radius	r_k_skv
Диаметр эксплуатационной колонны скважины Well production string diameter	ТР Двнутр ЭК [мм] TR Dinner PS [mm]
Инклинометрия Inclinometry	Смещение по линии устье–забой, м Displacement along the wellhead line, m
Назначение скважины Well purpose	Назнчачение_Нагнетательные, Назнчачение_Нефтяные, Назначение_Разведочные Appointment_Injection, Appointment_Oil, Appointment_Exploration
Плотность воды Water density	ТР Плотность воды [г/см³] TR Water density [g/cm³]
Плотность нефти Oil density	ТР Плотность нефти [г/см³] TR Oil density [g/cm³]
Коэффициент продуктивности Productivity factor	ТР К прод/TR Kprod
Пласт/Reservoir	Пласт_АС12, Пласт_АС9 Reservoir_AC12, Reservoir_AC9
Расстояние между скважинами Distance between wells	r_k_ryad
Плотность добываемой смеси Extracted mixture density	Плотность смеси расч [г/см³] Mixture density settlement [g/cm³]

Для обучения использовались следующие библиотеки:

- Модель леса деревьев. Этот метод относится к ансамблевым. Цель ансамблевых методов – объединить прогнозы нескольких базовых оценок, построенных с заданным алгоритмом обучения, чтобы улучшить надежность по сравнению с одной оценкой [17]. Суть данного метода заключается в том, что деревья обучаются не на всем наборе признаков, а на ограниченном. Далее, исходя из всех полученных оценок по разным деревьям, оценка усредняется и выдается итоговый ответ. Благодаря этому данная модель выигрывает у модели обучающего дерева, так как отсутствует переобучение.
- Модель обучающего дерева. Суть данного метода заключается в том, что каждая скважина проходит определенную цепочку условий, которые основаны на признаках, на которых обучается модель. В итоге, исходя из уникального для каждой скважины набора параметров, по ней выдается ответ: является она кандидатом или нет [18].
- Модель линейной регрессии. Суть данной модели состоит в том, что модель линеаризует набор признаков [19]. В итоге работы модели получается вероятность, с которой скважина будет являться кандидатом для проведения ОПЗ. Так, если эта вероятность составляет больше 0,5 (значение по умолчанию), скважина будет являться кандидатом для проведения ОПЗ.

При обучении модель может давать четыре типа ответов:

1. True Positive (TP). Это те ответы, на которые модель ответила положительно, и эти ответы действительно являются положительными.
2. True Negative (TN). Это те ответы, на которые модель ответила отрицательно, и эти ответы действительно являются отрицательными.
3. False Negative (FN). Это те ответы, на которые модель ответила отрицательно, но они на самом деле были положительными.
4. False Positive (FN). Это те ответы, на которые модель ответила положительно, но они на самом деле были отрицательными.

Исходя из этого для оценки сходимости результатов были разработаны две метрики. В соответствии с работой [16] первая из них, «precession», рассчитывалась по формуле (1):

$$precession = \frac{TP}{TP+FP}. \quad (1)$$

Данная метрика показывает долю объектов, названных алгоритмом положительными, и при этом действительно являющихся положительными.

В соответствии с работой [20] вторая метрика, «recall», рассчитывалась по формуле (2):

$$recall = \frac{TP}{TP+FN}. \quad (2)$$

Эта метрика показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Однако по отдельности данные метрики неэффективны, поэтому чаще всего в машинном обучении при задаче классификации используется метрика F1-score. В соответствии с работой [21] она рассчитывается по формуле (3):

$$f1 - score = \frac{2*precession*recall}{precession+recall}. \quad (3)$$

Чем ближе данная метрика к единице, тем точнее модель предсказывает ответ.

Также итоговая модель оценивалась при помощи метрики AUC-ROC. В соответствии с работой [21] площадь под ROC-кривой (обозначена синей заливкой) – это метрика оценки для задач бинарной классификации (рис. 1).

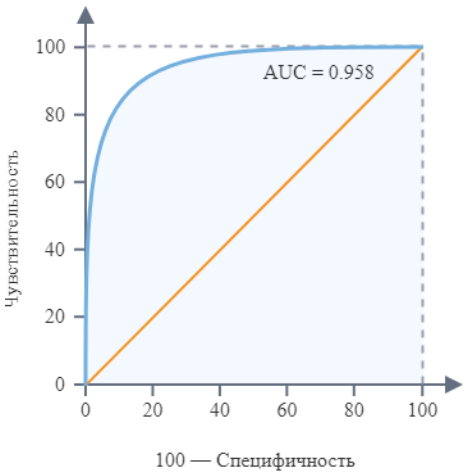

Рис. 1. *Метрика AUC-ROC*
Fig. 1. *AUC-ROC Metric*

На рис. 1 площадь под кривой (синяя линия) является мерой способности классификатора различать классы. Чем ближе значение площади к единице, тем лучше модель.

Для борьбы с дисбалансом использовались методы в соответствии с работой [17]:

1. *Threshold change*. Данный метод применялся для линейной регрессии. Модель линейной регрессии в итоге своей работы определяет вероятность, с которой скважина будет принадлежать тому или иному классу. Так, например, если вероятность равна 0,87, то скважина будет являться правильным кандидатом, так как это значение вероятности выше порогового, которое составляет 0,5 по умолчанию, если значение вероятности составляет, например, 0,3, то скважина не будет являться правильным кандидатом. В основе метода изменения порога лежит изменение значения по умолчанию.
2. *Upsampling*. Данный метод применялся для всех моделей. Его суть заключается в копировании строк того класса, который находится в меньшинстве. Данный метод не является предпочтительным, так как по своей сути он создает дубликаты, которые, как правило, мешают обучению модели.
3. *Downsampling*. Данный метод применялся для всех моделей. Его суть заключается в удалении строк того класса, который находится в большинстве. Данный метод является более предпочтительным, так как дубликаты не создаются. Однако могут быть упущены некоторые важные для обучения строки, поэтому применение данного метода также нежелательно.
4. *Balanced*. Данный метод применялся для всех моделей. Это встроенный в модель машинного обучения метод, который уменьшает влияние дисбаланса. Суть данного метода заключается в том, что смещаются веса классов данных в пропорции дисбаланса, за счет чего уменьшается важность превосходящего класса. Данный метод борьбы с дисбалансом классов совмещает в себе upsampling и downsampling. Он популярен в силу того, что нет необходимости вручную подбирать эти веса, в отличие от методов upsampling и downsampling.

Результаты

Осредненное значение метрики F1-score по кросс-валидации без использования методов борьбы с дисбалансом классов приведено в табл. 2.

Таблица 2. *Значение метрики F1_score без борьбы с дисбалансом*

Table 2. *Value of the F1_score metric without combating imbalance*

Модель Model	Значения метрики F1-score Value of the F1_score metric, %
Модель линейной регрессии LinearRegression	9,0
Модель леса деревьев RandomForestClassifier	96,5
Модель обучающего дерева RandomTreeClassifier	92,3

Из табл. 2 очевидно, что модель линейной регрессии не является корректной. При этом другие две модели показывают хорошую сходимость, несмотря на сильный дисбаланс классов.

Для лучшей сходимости и более корректной оценки работы моделей были применены следующие методы борьбы с дисбалансом классов: изменение порога значения определения скважин-кандидатов и скважин-некандидатов, метод upsampling, метод downsampling и метод balanced.

Результаты оценки сходимости моделей машинного обучения с реальными данными после применения различных способов борьбы с дисбалансом представлены в табл. 3.

Из табл. 3 видно, что лучше всего показала себя модель обучающего леса деревьев.

Таблица 3. *Значение метрики F1_score*

Table 3. *F1_score metric value*

Модель Model	Метод борьбы с дисбалансом Method of dealing with imbalance			
	Порог Threshold	Upsampling	Downsampling	Balanced
Модель линейной регрессии LinearRegression	11,5	9,1	10,9	5,3
Модель обучающего дерева RandomTreeClassifier	–	97,4	67,5	53,2
Модель леса деревьев RandomForestClassifier	–	99,0	94,7	99,1

Также необходимо отметить, что качество предсказания некоторых моделей упало при борьбе с дисбалансом классов.

1. При использовании методов upsampling и downsampling качество упало, так как над данными были проведены различные операции, которые ухудшают данные (создаются дубликаты, или, наоборот, удаляются нужные данные).
2. При использовании метода balanced качество для моделей линейной регрессии и обучающего дерева также упало в силу того, что дисбаланс оказывает слабое влияние на эти модели машинного обучения [22]. Поэтому борьба с дисбалансом может только ухудшить качество этих моделей.

Значение метрики F1-score является очень высоким, что говорит о корректности применения моделей машинного обучения для определения скважин-кандидатов для ОПЗ.

Далее для модели обучающего леса деревьев были подобраны наилучшие гиперпараметры методом GridSearchCV библиотеки sklearn в соответствии с работой [23].

На этих подобранных параметрах была обучена вся тренировочная выборка, и далее данная модель была проверена на тестовой выборке. Итоговая сходимость по метрике F1-score оказалась равна 99,5 %. При кросс-валидации тренировочная выборка делилась на четыре части, и одна часть использовалась для оценки сходимости. Для оценки сходимости на тестовой выборке вся тренировочная выборка использовалась для обучения. Поэтому возросла сходимость результатов при оценке модели на тестовой выборке.

Также помимо высокой точности следует отметить и скорость работы программы, которая содержит в себе модели машинного обучения. В целом программа обрабатывает 199 скважин анализируемого месторождения в течение 15 минут.

Также была оценена метрика AUC-ROC для наилучшей модели. Ее значение составило 99,9 %. Высокое значение этой метрики подтверждает, что модель обучающего леса способна на высоком уровне различать скважины-кандидаты и скважины-некандидаты. Это значит, что модель не «наугад» определяет скважины-кандидаты, соответственно, она намного лучше константной модели или модели, которая «наугад» отвечает, является ли скважина кандидатом.

Также дополнительно была построена константная модель для проверки модели леса деревьев на адекватность. Были созданы объекты с ответами, состоящие из одних нулей (все скважины являются некандидатами) и одних единиц (все скважины являются кандидатами). Были оценены сходимости данных константных моделей по метрике F1-score. Они составили 0,1 и 1 % соответственно. Это подтверждает адекватность построенной модели.

Исходя из табл. 3 можно сделать вывод о том, что наилучшей моделью машинного обучения для решения задачи классификации скважин-кандидатов является модель леса деревьев.

Подобранные гиперпараметры модели составили следующие величины: random_state=12345, n_estimators=30, max_depth=15, метод борьбы с дисбалансом классов – class_weight='balanced'. Гиперпараметры подбирались при помощи метода GridSearchCV, который автоматически находит наилучшие гиперпараметры для исследуемой метрики из указанного автором диапазона.

Также дополнительно было проведено исследование на важность признаков после обучения моделей (SHAP-анализ). Данное исследование проводилось при помощи функции feature_importances_ библиотеки sklearn в соответствии с работой [20, 24]. Гистограмма данного исследования представлена на рис. 2.

Анализируя рис. 2, можно сделать вывод о том, что наиболее важными признаками для обучения являются текущее забойное давление, текущее пластовое давление, дебит по жидкости, а также вязкость жидкости, остальные признаки имеют менее важную роль. Менее важными признаками являются: вязкость воды в пластовых условиях, расстояние между скважинами, а также внутренний диаметр эксплуатационной колонны. В будущем данные признаки будут удалены из модели машинного обучения для увеличения производительности модели.

Также следует отметить, что дебит газа был исключен из признаков для обучения, так как его значение сильно коррелирует со значением дебита жидкости, что может отрицательно повлиять на качество обучения модели.

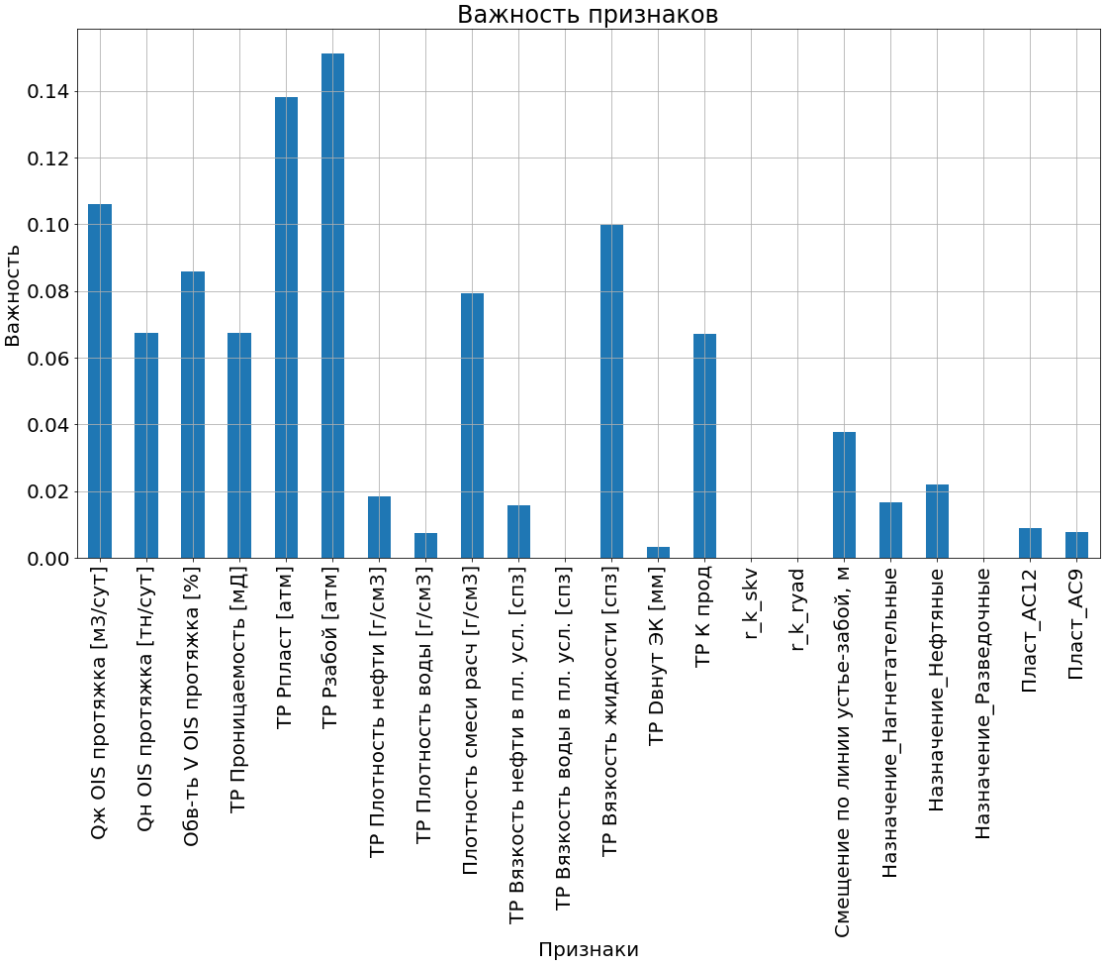

Рис. 2. *Исследование важности признаков*
Fig. 2. *Research of the importance of features*

Обсуждение

При исследовательском анализе данных, как было отмечено выше, было сделано допущение о том, что, если скважина объявляется кандидатом для проведения ОПЗ, она является кандидатом в течение всего месяца. Данное допущение было сделано в связи с дефицитом данных по скважинам-кандидатам из-за того, что месторождение было открыто недавно и находится только на второй стадии разработки. Также следует остановиться на допущении, связанном с тем, что ряд параметров, которые были перечислены выше, были заменены на один категорийный параметр: пласт. В связи с этим для применения этой модели на другом месторождении необходимо использовать другой набор признаков.

Также следует остановиться на основной выделенной в наборе данных проблеме: дисбаланс классов. Данная проблема была в итоге решена при помощи метода class_weight='balanced'. Следует отметить, что, в соответствии с работой [7], данный способ применяется довольно часто.

Проведенная работа показала, что использование моделей машинного обучения для предсказания ответа, будет ли скважина являться правильным кандидатом для проведения ОПЗ, является корректным, так как предсказания модели обладают высокой точностью. Модели также позволяют при помощи SHAP-анализа делать заключение о том, какие признаки были важны для обучения, а какие нет. Это позволяет впоследствии использовать лишь те признаки, которые улучшают качество модели для более корректной оценки.

Необходимо отметить, что данные, которые передаются модели для обучения, не всегда являются точными. Это происходит из-за того, что они получаются путем различных исследований, которые не всегда точны в силу различных технических причин. Поэтому точность предсказания модели и высокая сходимость все же не обеспечивает правильность выбора скважин-кандидатов. По этой причине необходимо повышать точность работы моделей. Для этого, а также для увеличения производительности модели планируется использование дру-

гих моделей машинного обучения, в том числе моделей градиентного бустинга и нейронных сетей. Также для повышения ценности моделей машинного обучения в производстве необходимо в том числе повышать точность и достоверность исходных измерений, которые подаются на вход алгоритму машинного обучения.

В ходе дальнейших исследований планируется улучшение качества и пользовательских характеристик программы путем создания удобного интерфейса и исключения из программы вышеперечисленных допущений.

Заключение

Проведенное исследование показало, что такая модель машинного обучения, как модель леса деревьев, позволяет предсказывать с точностью до 99,5 %, будет ли скважина являться правильным кандидатом для проведения ОПЗ. При этом указанная модель позволяет существенно сократить время обработки данных. Это говорит о перспективности применения указанной модели машинного обучения на реальных месторождениях для точного отбора скважин-кандидатов.

СПИСОК ЛИТЕРАТУРЫ

1. Кривощенков С.Н., Кочнев А.А., Равелев К.А. Разработка алгоритма определения технологических параметров нагнетания кислотного состава при обработке призабойной зоны пласта с учетом экономической эффективности // Записки Горного института. – 2021. – Т. 250. – С. 587–595.
2. Особенности формирования призабойных зон продуктивных пластов на месторождениях с высокой газонасыщенностью пластовой нефти / В.И. Галкин, Д.А. Мартюшев, И.Н. Пономарева, И.А. Черных // Записки Горного института. – 2021. – Т. 249. – С. 386–392.
3. Хасанов М.М., Мальцев А.А. Моделирование кислотной обработки полимиктового коллектора // Записки Горного института. – 2021. – Т. 251. – С. 678–687.
4. Рогачев М.К., Мухаметшин В.В. Контроль и регулирование процесса солянокислотого воздействия на призабойную зону скважин по геолого-промысловым данным // Записки Горного института. – 2018. – Т. 231. – С. 274–278.
5. Интенсификация притока нефти из карбонатных коллекторов для условий месторождений Западной Сибири / А.Д. Румянцев, А.М. Машкова, Н.В. Соловьев, К.О. Щербакова, Б.А. Овезов // Молодые – наукам о земле: Материалы X Международной научной конференции молодых ученых. – М.: Российский государственный геологоразведочный университет им. С. Орджоникидзе, 2022. – Т. 4. – С. 279–282.
6. Pessach D., Shmueli E. A review on fairness in machine learning // ACM Computing Surveys. – 2022. – Vol. 22. – № 3. URL: https://dl.acm.org/doi/10.1145/3494672 (дата обращения 19.09.2023).
7. Звездов А.В., Ерофеев В.А., Шираков В.Ю. Совершенствование технологии обработки призабойной зоны пласта заглинизированного терригенного коллектора // Инновационные технологии в нефтегазовой отрасли. Проблемы устойчивого развития территорий: Сборник трудов III Международной научно-практической конференции. – Ставрополь: Северо-Кавказский федеральный университет, 2022. – С. 251–258.
8. Гунькина Т.А. Методы повышения продуктивности нефтяных скважин. – Ставрополь: Северо-Кавказский Федеральный университет, 2017. – 57 с.
9. Применение машинного обучения для прогнозирования пластового давления при разработке нефтяных месторождений / Д.А. Мартюшев, И.Н. Пономарева, Л.А. Захаров, Т.А. Шадров // Известия Томского политехнического университета. Инжиниринг георесурсов. – 2021. – Т. 332. – № 10. – С. 140–149.
10. Разработка комплексной методики прогноза эффективности геолого-технических мероприятий на основе алгоритмов машинного обучения / А.А. Кочнев, Н.Д. Козырев, О.Е. Кочнева, С.В. Галкин // Георесурсы. – 2020. – Т. 22. – № 3. – С. 79–86.
11. Подбор скважин-кандидатов для гидравлического разрыва пласта на основе математического моделирования с использованием методов машинного обучения / А.Ф. Азбуханов, И.В. Костригин, К.А. Бондаренко, М.Н. Семенова, И.А. Середа, Д.Р. Юлмухаметов // Нефтяное хозяйство. – 2020. – № 11. – С. 38–42.
12. Курганов Д.В. Оценка эффективности обработок призабойных зон нефтяных скважин с применением методов машинного обучения // Автоматизация процессов управления. – 2020. – № 1 (59). – С. 47–54.
13. Федотова Л.Е. Предикативная аналитика результатов обработки призабойных зон пласта // Проблемы геологии и освоения недр: Труды XXIV Международного симпозиума имени академика М.А. Усова студентов и молодых учёных, посвященного 75-летию Победы в Великой Отечественной войне. – Томск: ТПУ, 2020. – Т. 2. – С. 155–156.
14. Machine learning with PyTorch and Scikit-learn: develop machine learning and deep learning models with Python / S. Raschka, Y. Liu, V. Mirjalili, D. Dzhulgakov. – Birmingham: Kindle Edition Publ., 2022. – 770 p.
15. Miranda F.M., Kohnecke N., Renard B.Y. Hiclass: a python library for local ierarchical classification compatible with scikit-learn // Journal of Machine Learning Research. – 2023. – № 24 (29). URL: https://www.jmlr.org/ (дата обращения 19.09.2023).
16. Yanli Liu, Yourong Wang, Jian Zhang. New Machine Learning Algorithm: Random Forest // Information Computing and Applications: Materials International Conference on Information Computing and Applications. – China: Springer Publ., 2012. – Vol. 7473. – P. 246–252.
17. An introduction to decision tree modeling / A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown // Journal of Chemometrics. – 2004. – Vol. 18. – P. 275–285. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/cem.873 (дата обращения 19.09.2023).
18. Hartmann F.G., Kopp J., Lois D. Social science data analysis. – Wiesbaden: Springer Publ., 2023. – 191 p.
19. Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth / A. Humphrey, W. Kuberski, J. Bialek, N. Perrakis, W. Cools, N. Nuyttens, H. Elakhrass, P.A.C. Cunha // Monthly Notes. – 2022. – Vol. 517. – P. 116–120. URL: https://academic.oup.com/mnrasl/article-abstract/517/1/L116/6761704 (дата обращения 19.09.2023).

20. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation / A.M. Carrington, D.G. Manuel, P.W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, A. Holzinger // National Library of Medicine. – 2023. – № 45 (1). – P. 329–341. URL: https://pubmed.ncbi.nlm.nih.gov/35077357/ (дата обращения 19.09.2023).
21. Stop oversampling for class imbalance learning: a review / A.S. Tarawneh, A.B. Hassanat, G.A. Altarawhen, A. Almuhaimmed // IEEE Access. – 2022. – Vol. 10. – P. 47643–47659.
22. Дале Д. Нужно ли бояться несбалансированности классов // Хабр. – 2018. URL: https://habr.com/ru/articles/349078/ (дата обращения 01.10.2023).
23. Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques / Z.M. Alhakeem, Y.M. Jebur, S.N. Henedy, H. Imran, L.F.A. Bernardo, H.M. Hussein // Materials. – 2022. – Vol. 2. – № 15 (21). URL: https://www.mdpi.com/1996-1944/15/21/7432 (дата обращения 19.09.2023).
24. On the role and the importance of features for background modeling and foreground detection / Th. Bouwmans, C. Silva, C. Marghes, M.S. Zitouni, H. Bhaskar, C. Frelicot // Computer Science Review. – 2018. – Vol. 28. – P. 26–91.

Информация об авторах

Максим Александрович Ямкин, студент, кафедра разработки и эксплуатации нефтяных и газовых месторождений, Санкт-Петербургский горный университет, Россия, 199106, г. Санкт-Петербург, 21-я лин. В.О., 2. makson.yamkin@mail.ru

Елена Улубековна Сафиуллина, кандидат технических наук, доцент, кафедра разработки и эксплуатации нефтяных и газовых месторождений, Санкт-Петербургский горный университет, Россия, 199106, г. Санкт-Петербург, 21-я лин. В.О., 2. safiullinaeu@yandex.ru

Александр Владимирович Ямкин, заместитель начальника технического отдела «ООО Газпром трансгаз Томск», Россия, 634029, г. Томск, пр. Фрунзе, 9. A.Yamkin@gtt.gazprom.ru

Поступила в редакцию: 11.10.2023
Поступила после рецензирования: 01.11.2023
Принята к публикации: 14.02.2024

REFERENCES

1. Krivoschenkov S.N., Kochnev A.A., Ravelev K.A. Development of an algorithm for determining the technological parameters of injection of an acid composition when treating the bottomhole formation zone, taking into account economic efficiency. *Notes of the Mining Institute*, 2021, vol. 250, pp. 587–595. (In Russ.)
2. Galkin V.I., Martyushev D.A., Ponomareva I.N., Chernykh I.A. Features of the formation of bottomhole zones of productive formations in fields with high gas saturation of reservoir oil. *Notes of the Mining Institute*, 2021, vol. 249, pp. 386–392. (In Russ.)
3. Khasanov M.M., Maltsev A.A. Simulation of acid treatment of a polymict reservoir. *Notes of the Mining Institute*, 2021, vol. 251, pp. 678–687. (In Russ.)
4. Rogachev M.K., Mukhametshin V.V. Control and regulation of the process of hydrochloric acid impact on the bottomhole zone of wells according to geological and field data *Notes of the Mining Institute*, 2021, vol. 231, pp. 274–278. (In Russ.)
5. Rumyantsev A.D., Mashkova A.M., Solovyev N.V., Shcherbakova K.O., Ovezov B.A. Intensification of oil inflow from carbonate reservoirs for the conditions of fields in Western Siberia. *Young people – to geosciences. Proc. of the X International Scientific Conference of Young Scientists*. Moscow, Russian State Geological Prospecting University named after. S. Ordzhonikidze Publ., 2022. Vol. 4, pp. 279–282. (In Russ.)
6. Pessach D., Shmueli E. A review on fairness in machine learning. *ACM Computing Surveys*, 2022, vol. 22, no. 3. Available at: https://dl.acm.org/doi/10.1145/3494672 (accessed 19 September 2023).
7. Zvezdov A.V., Erofeev V.A., Shirakov V.Yu. Improving the technology for processing the bottomhole zone of a clayey terrigenous reservoir. *Innovative technologies in the oil and gas industry. Problems of sustainable development of territories. Proc. of the III International Scientific and Practical Conference*. Stavropol, North Caucasus Federal University Publ., 2022. pp. 251–258. (In Russ.)
8. Gunkina T.A. *Methods for increasing oil well productivity*. Stavropol, North Caucasus Federal University Publ., 2017. 57 p. (In Russ.)
9. Martyushev D.A., Ponomareva I.N., Zakharov L.A., Shadrov T.A. Application of machine learning to predict reservoir pressure in oil field development. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering*, 2021, vol. 332, no. 10, pp. 140–149. (In Russ.)
10. Kochnev A.A., Kozyrev N.D., Kochneva O.E., Galkin S.V. Development of a comprehensive methodology for predicting the effectiveness of geological and technical activities based on machine learning algorithms. *Georesources*, 2020, vol. 22, no. 3, pp. 79–86. (In Russ.)
11. Azbukhanov A.F., Kostrigin I.V., Bondarenko K.A., Semenova M.N., Sereda I.A., Yulmukhametov D.R. Selection of candidate wells for hydraulic fracturing based on mathematical modeling using machine learning methods. *Oil industry*, 2020, no. 11, pp. 38–42. (In Russ.)
12. Kurganov D.V. Assessing the effectiveness of treatments for bottom-hole zones of oil wells using machine learning methods. *Automation of management processes*, 2020, no. 1 (59), pp. 47–54. (In Russ.)
13. Fedotova L.E. Predictive analytics of the results of processing near-wellbore formation zones. *Problems of geology and subsoil development. Proc. of the XXIV International Symposium named after Academician M.A. Usov of students and young scientists,*

dedicated to the 75th anniversary of Victory in the Great Patriotic War. Tomsk, TPU Publ., 2020. Vol. 2, pp. 155–156. (In Russ.)
14. Raschka S., Liu Y., Mirjalili V., Dzhulgakov D. *Machine Learning with PyTorch and Scikit-learn: develop machine learning and deep learning models with Python*. Birmingham, Kindle Edition Publ., 2022. 770 p.
15. Miranda F.M., Kohnecke N., Renard B.Y. Hiclass: a python library for local ierarchical classification compatible with scikit-learn. *Journal of Machine Learning Research*, 2023, no. 24 (29). Available at: https://www.jmlr.org/ (accessed 19 September 2023).
16. Yanli Liu, Yourong Wang, Jian Zhang. New Machine Learning Algorithm: Random Forest. *Information Computing and Applications. Materials International Conference on Information Computing and Applications*. China, Springer Publ., 2012. Vol. 7473, pp. 246–252.
17. Myles A.J., Feudale R.N., Liu Y., Woody N.A., Brown S.D. An introduction to decision tree modeling. *Journal of Chemometrics*, 2004, vol. 18, pp. 275–285. Available at: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/cem.873 (accessed 19 September 2023).
18. Hartmann F.G., Kopp J., Lois D. *Social Science Data Analysis*. Wiesbaden, Springer Publ., 2023. 191 p.
19. Humphrey A., Kuberski W., Bialek J., Perrakis N., Cools W., Nuyttens N., Elakhrass H., Cunha P.A.C. Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth. *Monthly Notes*, 2022, vol. 517, pp. 116–120. Available at: https://academic.oup.com/mnrasl/article-abstract/517/1/L116/6761704 (accessed 19 September 2023).
20. Carrington A.M., Manuel D.G., Fieguth P.W., Ramsay T., Osmani V., Wernly B., Bennett C., Hawken S., Magwood O., Sheikh Y., McInnes M., Holzinger A. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *National Library of Medicine*, 2023, no. 45 (1), pp. 329–341. Available at: https://pubmed.ncbi.nlm.nih.gov/35077357/ (accessed 19 September 2023).
21. Tarawneh A.S., Hassanat A.B., Altarawhen G.A. Stop oversampling for class imbalance learning: a review. *IEEE Access*, 2022, vol. 10, pp. 47643–47659.
22. Dale D. Should we be afraid of class imbalance. *Habr*, 2018. Available at: https://habr.com/ru/articles/349078/ (accessed 1 October 2023).
23. Alhakeem Z.M., Jebur Y.M., Henedy S.N., Imran H., Bernardo L.F.A., Hussein H.M. Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques. *Materials*, 2022, vol. 2, no. 15 (21). Available at: https://www.mdpi.com/1996-1944/15/21/7432 (accessed 19 September 2023).
24. Bouwmans Th., Silva C., Marghes C., Zitouni M.S., Bhaskar H., Frelicot C. On the role and the importance of features for background modeling and foreground detection. *Computer Science Review*, 2018, vol. 28, pp. 26–91.

Information about the authors

Maxim A. Yamkin, Student, St. Petersburg Mining University, 2, 21st line V.O., St. Petersburg, 199106, Russian Federation. makson.yamkin@mail.ru
Elena U. Safiullina, Cand. Sc., Associate Professor, St. Petersburg Mining University, 2, 21st line V.O., St. Petersburg, 199106, Russian Federation. safiullinaeu@yandex.ru
Alexander V. Yamkin, Deputy Head of Technical Department, «Gazprom transgaz Tomsk LLC», 9, Frunze avenue, Tomsk, 634029, Russian Federation. A.Yamkin@gtt.gazprom.ru

Received: 11.10.2023
Revised: 01.11.2023
Accepted: 14.02.2024