ЦИФРОВЫЕ ТЕХНОЛОГИИ

УДК 622.245:004.438Python
https://doi.org/10.51890/2587-7399-2024-9-4-152-158

ПРЕДОБРАБОТКА ДАННЫХ ДЛЯ ИНТЕГРИРОВАННОГО МОДЕЛИРОВАНИЯ АКТИВОВ

© Коллектив авторов, 2024

CC BY 4.0

К.А. Печко[1,*], Д.И. Константинов[2], М.В. Симонов[3], А.А. Афанасьев[3]
[1]Ассоциация «Цифровые технологии в промышленности», РФ, Санкт-Петербург
[2]Научно-образовательный центр «Газпромнефть-СГУ», РФ, Санкт-Петербург
[3]Группа компаний «Газпром нефть», РФ, Санкт-Петербург

Электронный адрес: ProNeft@gazprom-neft.ru

Введение. В настоящее время все активнее применяется машинное обучение во всех сферах производства, в том числе и в нефтяной промышленности. Однако качество данных, поступающих с нефтяных промыслов, не всегда позволяет корректно использовать их в процессах цифровизации производственных процессов. Быстрое переоборудование всех месторождений для точного и частого сбора данных затруднено. В связи с этим все же приходится работать с уже собранными данными.

Цель. Рассмотрение совокупности преобразований, которые происходят с данными в процессе их использования в машинном обучении, как единого процесса (ETL-процесса).

Материалы и методы. В качестве примеров для демонстрации рассматриваемых проблем и подходов использованы данные об эксплуатации нефтяных скважин. Для анализа и визуализации данных реализованы скрипты на языке программирования Python.

Результаты. В результате проведенной работы установлено, что качество данных, поступающих с нефтяных промыслов, не всегда позволяет использовать их в машинном обучении. Для повышения качества данных на этапе сбора и их подготовки предлагается использовать технологию ETL-процессов.

Заключение. Применение ETL-процессов позволит значительно увеличить количество и качество данных, которые можно использовать для создания цифровых двойников месторождений. Таким образом, сложно переоценить эффект от внедрения рассматриваемой технологии.

Ключевые слова: Extract Transform Load (ETL), Exploration Data Analysis (EDA), предобработка данных, машинное обучение, анализ данных, моделирование, Интегрированное Моделирование Активов (ИМА)

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Печко К.А., Константинов Д.И., Симонов М.В., Афанасьев А.А. Предобработка данных для интегрированного моделирования активов. PROНЕФТЬ. Профессионально о нефти. 2024;9(4):152–158. https://doi.org/10.51890/2587-7399-2024-9-4-152-158

Статья поступила в редакцию 02.07.2024
Принята к публикации 16.08.2024
Опубликована 27.12.2024

DATA PREPROCESSING FOR INTEGRATED ASSET MODELING

Konstantin A. Pechko[1,*], Dmitry I. Konstantinov[2], Maksim V. Simonov[3], Aleksandr A. Afanasev[3]
[1]Digital Technologies in Industry, RF, Saint Petersburg
[2]Science Education Center “Gazpromneft SGU”, RF, Saint Petersburg
[3]Gazprom neft company group, RF, Saint Petersburg

E-mail: ProNeft@gazprom-neft.ru

Introduction. Machine learning is increasingly being applied in various industries, including the oil and gas sector. However, the quality of data collected from oilfields does not always allow for its correct use in the digitalization of production processes. Unfortunately, it is not possible to quickly re-equip all oilfields for more accurate and frequent data collection. As a result, we must continue working with the data that has already been collected.

Objective. The objective of this work is to examine the set of transformations that occur with data in the process of its use in machine learning as a unified process (the ETL process).

Materials and methods. As examples to demonstrate the discussed challenges and approaches, we used data on oil well operations. Python scripts were developed for data analysis and visualization.

Results. The study found that the quality of data collected from oilfields is not always sufficient for use in machine learning. To improve data quality during collection and preparation stages, it is proposed to implement ETL processes.

Conclusion. The application of ETL processes will significantly increase the quantity and quality of data available for creating digital twins of oilfields. Therefore, the impact of introducing this technology is difficult to overestimate.

Keywords: Extract Transform Load (ETL), Exploration Data Analysis (EDA), data preprocessing, machine learning, data analysis, modeling, Integrated Asset Modeling (IAM)

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

For citation: Pechko K.A., Konstantinov D.I., Simonov M.V., Afanasev A.A. Data preprocessing for integrated asset modeling. PRONEFT. Professionally about oil. 2024;9(4):152–158. https://doi.org/10.51890/2587-7399-2024-9-4-152-158

Manuscript received 02.07.2024
Accepted 16.08.2024
Published 27.12.2024

ВВЕДЕНИЕ

В настоящее время методы машинного обучения (МО) находят свое применение во все большем количестве областей. Не осталась в стороне и нефтегазовая отрасль. Алгоритмы машинного обучения активно внедряются в программные продукты, применяемые на всех стадия разведки, добычи, транспортировки и переработки нефти и газа. Например, в данный момент идет работа над созданием модуля для расчета Vertical Lift Performance (VLP) кривых, в основе которого лежат обучающиеся алгоритмы.
Применение МО подразумевает наличие некоторых априорных (исторических) данных. В случае скважин речь идет о данных эксплуатации скважины: дебит по жидкости, обводненность добываемой продукции, газовый фактор, устьевое и забойное давление.
Прежде всего, необходимо провести тщательный анализ данных. Важно убедиться, что предоставленные данные соответствуют формату, который может быть воспринят моделью. Также требуется проверить адекватность значений с физической точки зрения, определить распределение данных и другие характеристики. Иными словами, необходимо изучить данные, с которыми придется взаимодействовать. Это исследование носит название разведочного анализа данных (Exploration Data Analysis, EDA) [3].
Как будет показано в данной статье, зачастую, после проведения EDA, выясняется, что данные имеют весьма низкое качество, и модель, построенная на таких данных, будет иметь низкую предсказательную способность. Соответственно, перед использованием данных, необходимо выполнить их дополнительную обработку (провести препроцессинг данных). А так как данные постоянно обновляются, а значит и модели тоже, и характер факторов, снижающих качество данных, практически не меняется с течением времени, то логично автоматизировать процесс предобработки данных [6]. Для этого в цепочку создания моделей необходимо встроить так называемый ETL-процесс.

EDA (EXPLORATORY DATA ANALYSIS)

Как уже было сказано выше, EDA — один из первых этапов в процессе создания алгоритма на основе МО. Целью данного этапа является погружение в имеющиеся данные, изучение их структуры и состава, выявление закономерностей в данных, определение их статистических характеристик, выявление аномалий, выдвижение и проверка гипотез.
Пример выполнения EDA для данных технологического режима скважин некоторой группы месторождений:

Изучение структуры данных
Для изучения структуры данных зачастую достаточно визуализировать их в исходном состоянии. Так как табличный формат представления данных является одним из самых удобных для анализа и обработки, то предпочтительно сводить исходные данные именно к этому формату [5]. Далее в статье подразумевается, что все данные представлены именно в табличном виде. В таком случае можно определить количество строк и столбцов, значение, тип данных.
Так, из анализа типичной ШТР (шахматка технологического режима) за 1 год по 150–160 скважинам видно, что она содержит порядка 40 тысяч строк и 30 столбцов. Данные в основном представлены вещественными числами, реже встречаются целочисленные значения, строковые данные и даты.

ДЛЯ СОЗДАНИЯ ЦИФРОВЫХ ДВОЙНИКОВ МЕСТОРОЖДЕНИЙ ЧАСТО ПРИМЕНЯЕТСЯ МАШИННОЕ ОБУЧЕНИЕ. ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ДАННЫХ, ПОСТУПАЮЩИХ С НЕФТЯНЫХ ПРОМЫСЛОВ, С ЦЕЛЬЮ ИСПОЛЬЗОВАНИЯ ИХ В МАШИННОМ ОБУЧЕНИИ НА ЭТАПЕ СБОРА И ИХ ПОДГОТОВКИ ПРЕДЛАГАЕТСЯ ИСПОЛЬЗОВАТЬ ТЕХНОЛОГИЮ ETL-ПРОЦЕССОВ.

Выявление закономерностей и взаимосвязей в данных
Самый простой метод выявления зависимостей в данных — графический. В процессе EDA в принципе часто приходится строить различного рода диаграммы и графики. Так, по диаграмме рассеяния можно определить вид зависимости двух параметров между собой.
На **рис. 1** представлен пример диаграммы рассеяния для данных по одной из скважин. На левом рисунке представлена диаграмма рассеяния для дебита жидкости и забойного

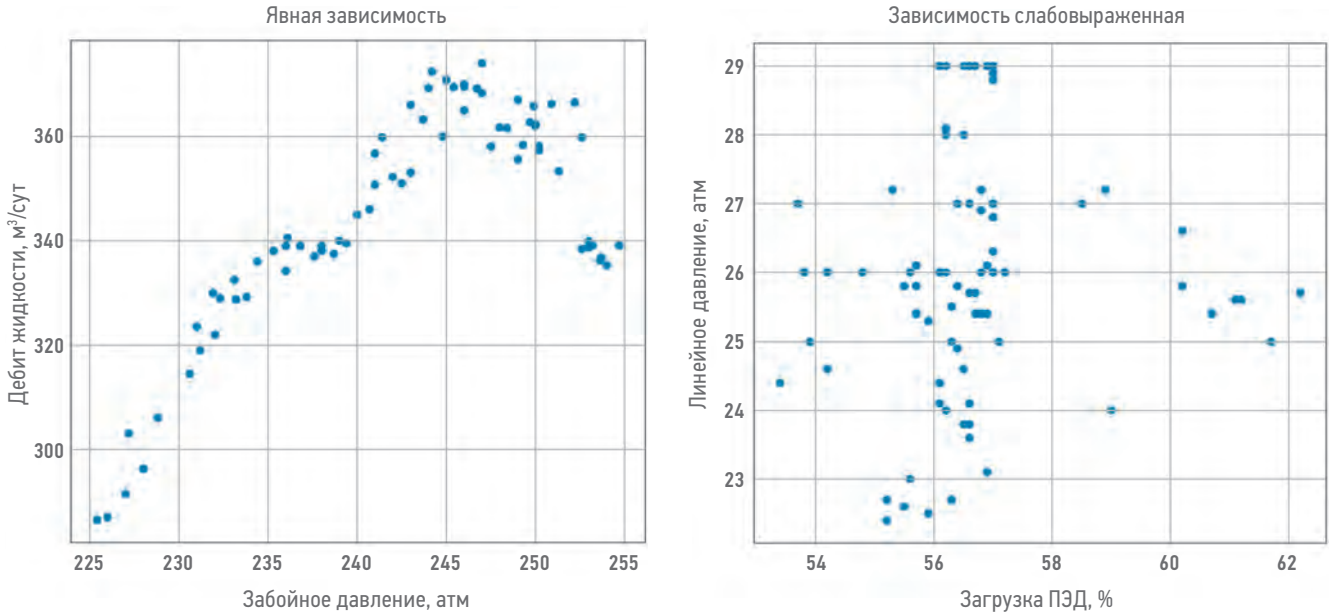

Рис. 1. Примеры зависимостей данных для месторождения X. Нарисовано при помощи Python (Д.И. Константинов)
Fig. 1. Examples of data dependencies for field X. Drawn using Python (Dmitry I. Konstantinov)

давления, здесь явно видна зависимость между этими параметрами. На правом рисунке очевидной зависимости не наблюдается, из чего можно сделать вывод, что загрузка погружного электродвигателя (ПЭД-а) и линейное давление на скважине не связаны между собой, по крайней мере линейно.

Статистический анализ данных

Статистический анализ — совокупность методов, направленных на определение закономерностей и основных тенденций в распределении данных. В частности, на этом этапе определяются такие статистические параметры как среднее значение, мода, медиана, дисперсия, а также определяется вид распределения данных (нормальное, пуассоновское, биномиальное и др.) [2].

На **рис. 2** представлен пример распределения значений основных параметров, характеризующих работу скважины. Из представленных гистограмм видно, что распределение данных в основном мультимодально, асимметрично и не является нормальным. Такая ситуация типична для нефтегазовой отрасли и является затрудняющим фактором при обработке данных и построении моделей [1].

Выдвижение гипотез и их проверка

Поскольку анализируемые данные в дальнейшем применяются для решения бизнес-кейсов, по завершении анализа выдвигаются гипотезы о наиболее эффективных способах их использования, значимости отдельных параметров и взаимосвязях между ними.

Так, специфика задачи требует использования только пяти признаков для обучения модели (буферное давление, забойное давление, дебит жидкости, обводненность и газовый фактор, т.к. задача модели — по давлению на устье скважины и по составу потока в трубе предсказать давление на забое скважины), но в процессе корреляционного анализ выяснилось, что значительно лучше связь забойного давления с дебитом газа, а не с газовым фактором. После испытания моделей на новом наборе данных было подтверждено улучшение качества моделей.

ETL

ETL (Extract, Transform, Load) — процесс переноса данных из одного хранилища (Extract — извлечение) в другое (Load — загрузка), в процессе которого данные также зачастую приводятся к требуемому формату (Transform — преобразование) [4].

В контексте рассматриваемой задачи (создание моделей скважин) этот процесс состоит из следующих этапов:

1) извлечение данных — данные о параметрах эксплуатации скважин хранятся в разных форматах в зависимости от требований добывающих компаний, соответственно нужно разработать систему, которая способна извлекать требуемые данные из разных форматов;

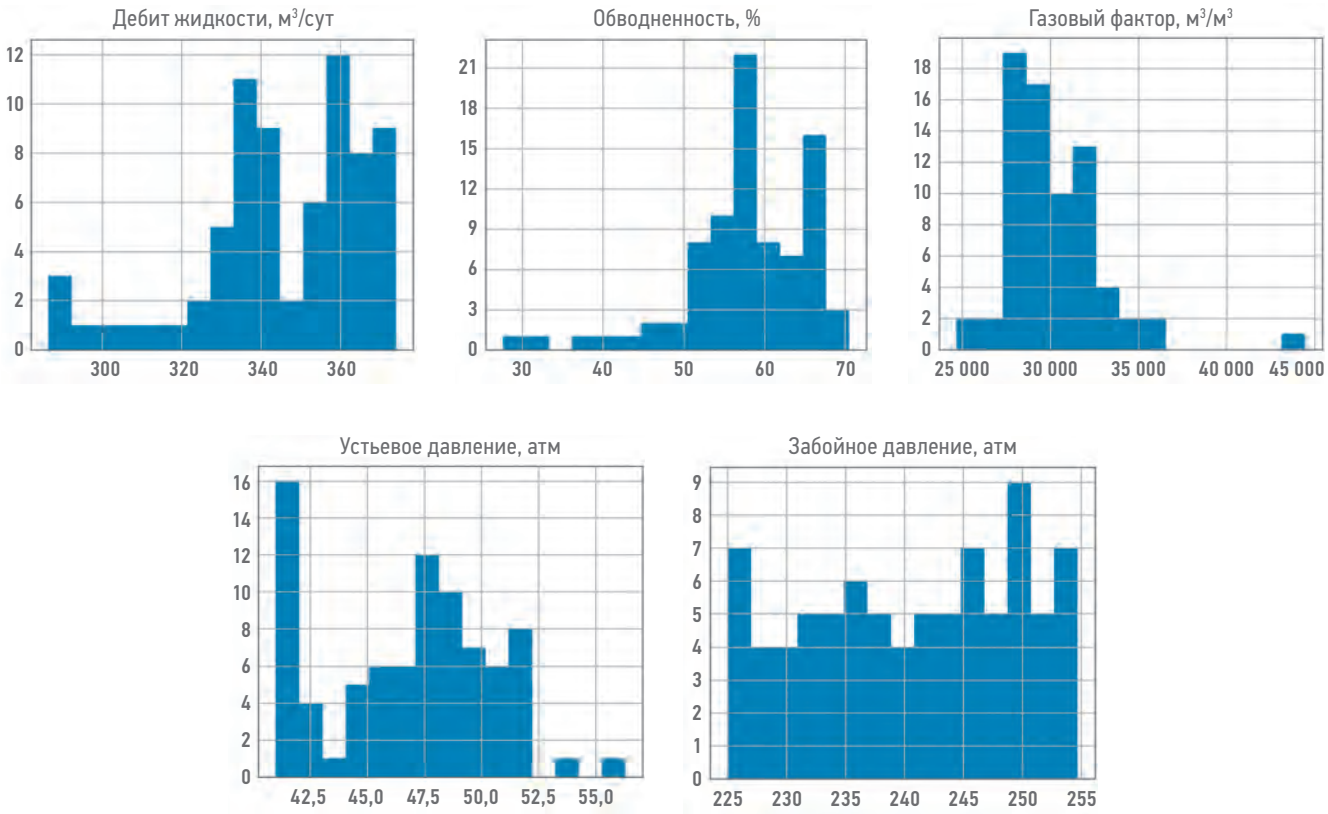

Рис. 2. Пример реальных гистограмм величин, используемых для построения интегрированных моделей для месторождения X. Нарисовано при помощи Python (Д.И. Константинов)
Fig. 2. An example of real histograms of values used to build integrated asset models for field X. Drawn using Python (Dmitry I. Konstantinov)

2) трансформация данных — после извлечения данных их необходимо подготовить к дальнейшему использованию, например необходимо корректно обработать пропуски в данных, привести данные к одному масштабу, отфильтровать выбросы, иногда нужно привести данные к какому-то конкретному виду распределения (чаще всего к нормальному) и т.д. В общем необходимо повысить качество данных;
3) загрузка данных — сохранение предобработанных данных в некоторой системе хранения в формате, удобном для использования.

Первый и последний шаги не так интересны с точки зрения анализа данных, поэтому им не будет уделено внимание в данной статье. В то же время шаг по трансформации данных не так прост, как может показаться. Основные особенности и сложности этого шага и будут рассмотрены далее более детально.

TRANSFORM

Специфика применения алгоритмов МО диктует два основных требования к данным, подающимся на вход алгоритму: достаточный их объем и удовлетворительное качество этих данных. Однако, далеко не всегда эти требования выполняются. Почти всегда данных недостаточно, а качество тех данных, которые все же имеются, находится на низком уровне.

Этап преобразования данных призван минимизировать указанные проблемы. Для этого весь процесс разбит на ряд задач, решая которое шаг за шагом, удается повысить качество данных. Данную последовательность задач можно представить в виде конвейера обработки данных [8], схема которого показана на **рис. 3**.

Рассмотрим эти задачи.

Очистка данных

Очень часто в данных встречаются так называемые выбросы — данные, значение которых сильно отклоняется от остальных. Реальный пример приведен на **рис. 4**. Как видно, одно из значений дебита явно выделяется на фоне остальных. Скорее всего, такой выброс связан с каким-то сбоем в измерительном оборудовании. Такое значение не отражает реальную работу скважины и оно не должно попасть в обучающую выборку.

Для фильтрации таких данных можно использовать различные методы. Например, правило трех сигм, интерквартильный

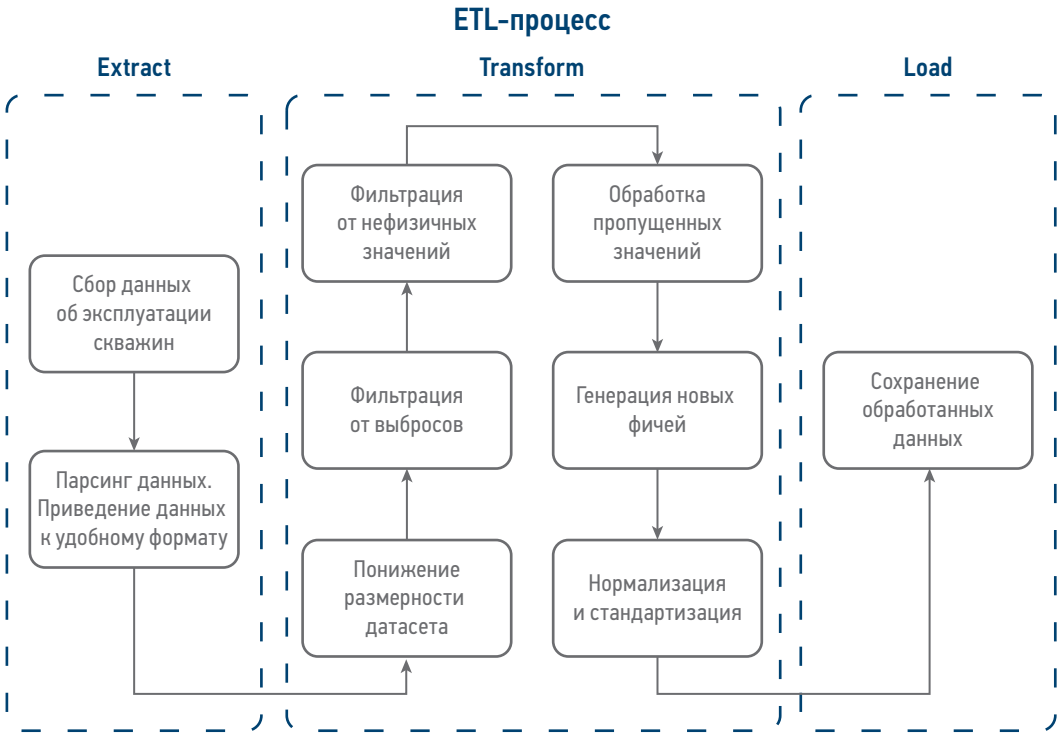

Рис. 3. Схема ETL процесса в ИМА с применением машинного обучения (Д.И. Константинов)
Fig. 3. Scheme of ETL process in IMA using machine learning (Dmitry I. Konstantinov)

размах и др. При этом правило фильтрации необходимо выбирать осторожно, ведь каждое отфильтрованное значение уменьшает размер обучающей выборки, а в условиях дефицита данных это становится критичной проблемой.

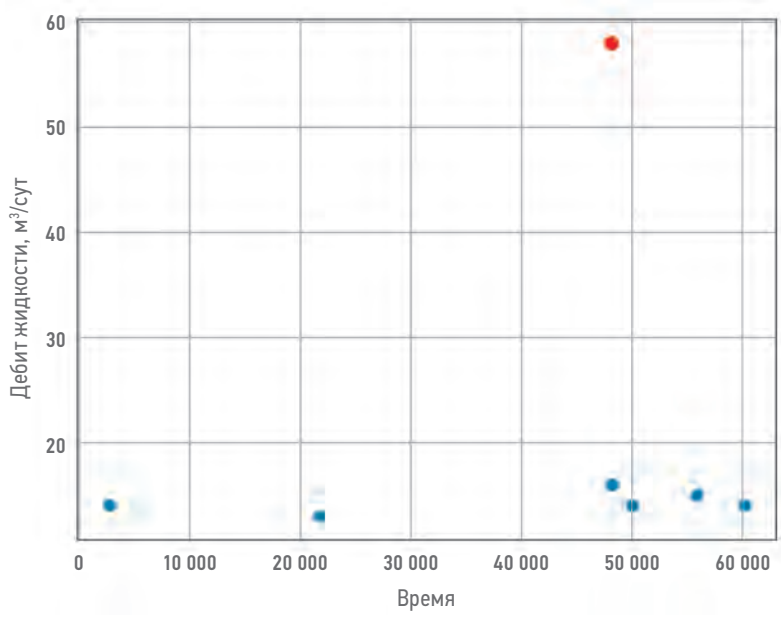

Рис. 4. Пример выброса в реальных данных. Нарисовано при помощи Python (Д.И. Константинов)
Fig. 4. Example of an outlier in real-world data. Drawn using Python (Dmitry I. Konstantinov)

Обработка пропущенных значений

Если посмотреть на распределение количества строк данных, содержащих различное количество пропусков (**рис. 5**), то можно увидеть, что только чуть менее 15 % строк не имеет пропусков, а в остальных строках имеется хотя бы один пропуск. А если посмотреть на то, какие данные чаще всего пропущены (**рис. 6**), то видно, что в основном имеются проблемы с замерами обводненности продукции.

Для качественного обучения модели, пропусков в данных быть не должно, но если просто удалять строки с пропущенными значениями, то снова возникает проблема уменьшения обучающей выборки [7]. Следовательно, необходимо разработать алгоритм, позволяющий восстанавливать хотя бы часть данных. Однако следует учесть вероятность, что в данных появятся новые зависимости, которых не было до этого.

В частности, имеется предположение, что можно пытаться восстановить строки, в которых отсутствует только один фактор (а таких строк большинство), без особого риска потерять истинную зависимость параметров и таргета. В будущем планируется более детально провести анализ этой проблемы и провести сравнительные тесты.

Feature engineering

Иногда для лучшей работы алгоритмов МО, к обучающим данным стоит добавлять новые

поля, которых не было в изначальном наборе. Например, зная дебит жидкости и обводненность, можно рассчитать дебит нефти или добавить некоторые полиномиальные параметры (например квадраты и попарные произведения всех параметров). Последнее может пригодится в ситуации, описанной ниже.

Большая часть алгоритмов, используемых при построении моделей, результатом возвращают некоторую кусочно-постоянную функцию, что является проблемой при дальнейшем использовании такого результата. Так алгоритм решения узлового анализа, который работает с применением градиентных методов, перестает корректно работать на участках с постоянным значением функции.

Можно использовать алгоритмы, дающие непрерывные функции, но их производительность и точность, зачастую, сильно ниже, что ограничивает возможности по их применению. Но, добавив к данным полиномиальные члены, обычная линейная регрессия начинает вести себя уже не линейно, имея при этом хорошую производительность.

Нормализация и стандартизация

Некоторые алгоритмы МО лучше работают с данными, которые имеют определенный формат. Например, крайне желательно, чтобы все данные были приведены к одному диапазону (обычно [0,1]). Для этого применяется нормализация данных.

Но и нормализации данных не всегда достаточно. Как было показано выше, данные о работе скважины зачастую имеют распределение далекое от нормального. При этом часть алгоритмов (в особенности на основе нейронных сетей) требуют нормальности распределения данных для корректной работы. Для приведения распределения данных к нормальному виду применяются различные методы стандартизации.

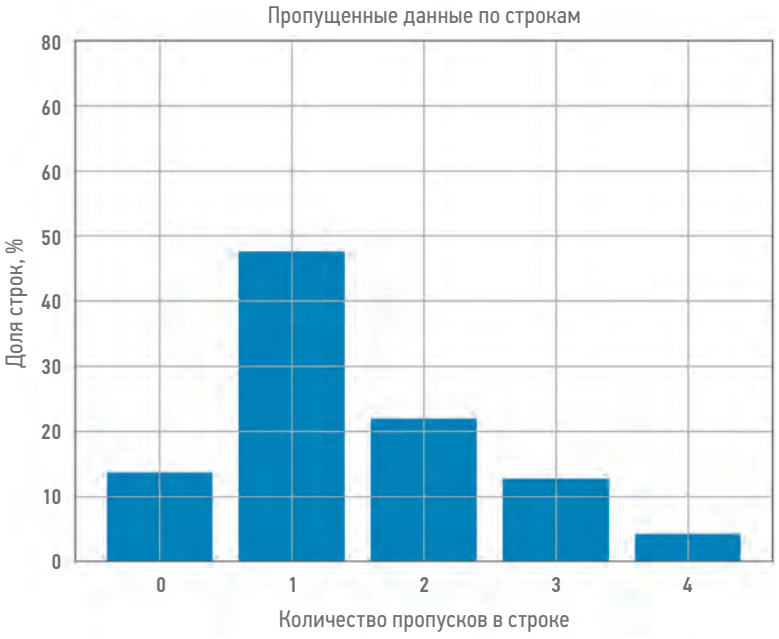

Рис. 5. Диаграмма количества пропущенных данных в наблюдении на одну дату для месторождения X. Нарисовано при помощи Python (Д. И. Константинов)
Fig. 5. Diagram of the number of missing data points per observation date for field X. Drawn using Python (Dmitry I. Konstantinov)

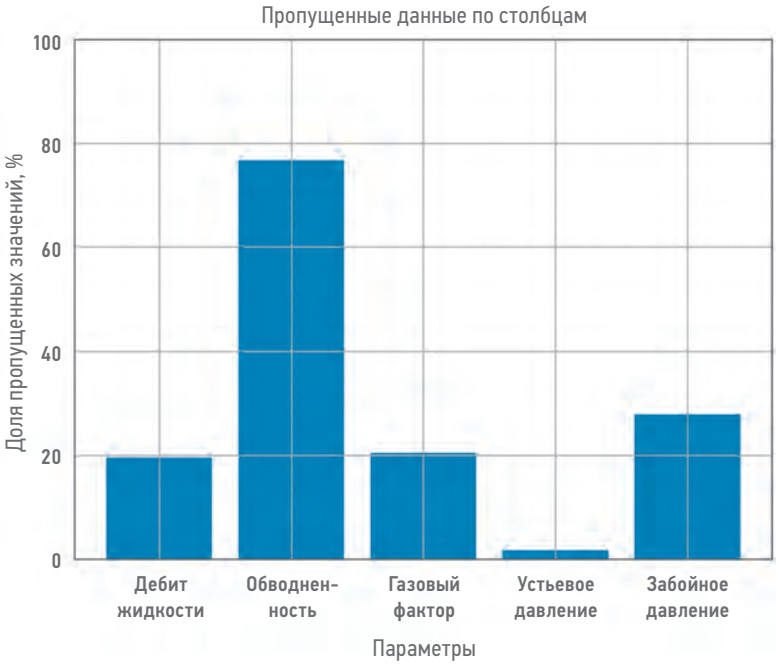

Рис. 6. Диаграмма типа пропущенных данных в наблюдении на одну дату для месторождения X. Нарисовано при помощи Python (Д. И. Константинов)
Fig. 6. Diagram of the types of missing data in observations on a single date for field X. Drawn using Python (Dmitry I. Konstantinov)

ЗАКЛЮЧЕНИЕ

В настоящее время все активнее применяется машинное обучение во всех сферах производства, в том числе и в нефтяной промышленности. Однако крайне низкое качество данных, поступающих с промыслов, является серьезным препятствием на пути цифровизации. Быстрое переоборудование всех месторождений для точного и частого сбора данных затруднено. В связи с этим все же приходится работать с уже имеющимися данными. Применение ETL-процессов позволит значительно увеличить количество и качество данных, которые можно использовать для создания цифровых двойников месторождений. Таким образом, сложно переоценить эффект от внедрения рассматриваемой технологии.

Список литературы
1. *Печко К.А., Сенькин И.С., Белоногов Е.В.* Моделирование скважин методами машинного обучения для задач интегрированного моделирования // PROНЕФТЬ. Профессионально о нефти. — 2022. — № 7(2). — С. 114–120. https://doi.org/10.51890/2587-7399-2022-7-2-114-120
2. *Fan C. et al.* A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in energy research. 2021, vol. 9, p. 652801.
3. García S. et al. Big data preprocessing: methods and prospects. Big data analytics. 2016, vol. 1, pp. 1–22.
4. *Malik J.S., Goyal P., Sharma A.K.* A comprehensive approach towards data preprocessing techniques & association rules. Proceedings of the 4th National Conference. 2010, vol. 132.
5. *Katser I.D. et al.* Data pre-processing methods for NPP equipment diagnostics algorithms: an overview. Nuclear Energy and Technology. 2021, vol. 7, no. 2, pp. 111–125.
6. *Быков К.В.* Особенности предобработки данных для применения машинного обучения // Молодой ученый. — 2021. — № 53 (395). — С. 1.
7. *Печко К.А., Чупров А.А., Афанасьев А.А., Симонов М.В.* Интерполяция пропусков данных технологического режима скважин алгоритмами машинного обучения // PROнефть. Профессионально о нефти. — 2023. — Т. 8. — № 3 — С. 163–166. https://doi. org/10.51890/2587-7399-2023-8-3-163-166
8. *Pechko K. et al.* Data Pre Processing Techniques in Integrated Asset Modeling // Fourth EAGE Digitalization Conference & Exhibition. *European Association of Geoscientists & Engineers*. 2024, vol. 2024, no. 1, pp. 1–4. (In Russ.)

References
1. Pechko K.A., Senkin I.S., Belonogov E.V. Well modeling using machine learning methods for integrated modeling. *PROneft. Professionally about Oil*. 2022, no. 7(2), pp. 114–120. (In Russ.) https://doi.org/10.51890/2587-7399-2022-7-2-114-120
2. Fan C. et al. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in energy research. 2021, vol. 9, p. 652801.
3. García S. et al. Big data preprocessing: methods and prospects. *Big data analytics*. 2016, vol. 1, pp. 1–22.
4. Malik J.S., Goyal P., Sharma A.K. *A comprehensive approach towards data preprocessing techniques & association rules.* Proceedings of the 4th National Conference. 2010, vol. 132.
5. Katser I.D. et al. Data pre-processing methods for NPP equipment diagnostics algorithms: an overview. *Nuclear Energy and Technology*. 2021, vol. 7, no. 2, pp. 111–125.
6. Bykov K. V. Specifics of data preprocessing for the application of machine learning. *Young scientist.* 2021, no. 53 (395), p. 1. (In Russ.)
7. Pechko K.A., Chuprov A.A., Afanasev A.A., Simonov M.V. Well data preprocessing using machine learning algorithms. *PROneft. Professionally about Oil.* 2023, no. 8(3), pp. 163–166. (In Russ.). https://doi.org/10.51890/2587-7399-2023-8-3-163-166
8. Pechko K. et al. Data Pre Processing Techniques in Integrated Asset Modeling // Fourth EAGE Digitalization Conference & Exhibition. *European Association of Geoscientists & Engineers*. 2024, vol. 2024, no. 1, pp. 1–4. (In Russ.)

ВКЛАД АВТОРОВ / AUTHOR CONTRIBUTIONS

К.А. Печко — разработал концепцию статьи, активно участвовал в подготовке текста статьи, в тестировании алгоритмов, оказывал экспертную поддержку.
Д.И. Константинов — подготовил текст статьи, реализовал программную часть работ, подготовил визуализацию примеров.
А.А. Афанасьев — оказал экспертную поддержку в области интегрированного моделирования, окончательно утвердил публикуемую версию статьи.
М.В. Симонов — руководил работой, контролировал качество и сроки выполняемых работ.

Konstantin A. Pechko — developed the article concept, took an active part in the preparation of the text and provided expert support.
Dmitriy I. Konstantinov — prepared the text of the article, implemented the software part of the work, prepared visualization of the samples.
Aleksandr A. Afanasev — provided expert support in integrated modeling, approved the final version of the article.
Maksim V. Simonov — work management, quality control and timing of work performed.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Константин Анатольевич Печко* — главный аналитик данных, Ассоциация «Цифровые технологии в промышленности»
190000, Россия, г. Санкт-Петербург, ул. Гороховая, д. 16, А5.
e-mail: ProNeft@gazprom-neft.ru
Scopus: 57331243400
ORCID: https://orcid.org/0000-0002-9663-8660

Дмитрий Игоревич Константинов — специалист НОЦ «Газпромнефть СГУ»

Александр Андриянович Афанасьев — главный специалист, Группа компаний «Газпром нефть»

Максим Владимирович Симонов — руководитель центра, Группа компаний «Газпром нефть»

Konstantin A. Pechko* — Chief data analytics officer, Digital Technologies in Industry
16, Gorohovaya str., 190000, Saint Petersburg, Russia.
e-mail: ProNeft@gazprom-neft.ru
Scopus: 57331243400
ORCID: https://orcid.org/0000-0002-9663-8660

Dmitriy I. Konstantinov — Specialist Science Education Center "Gazpromneft SGU"

Aleksandr A. Afanasev — Chief specialist, Gazprom neft company group

Maksim V. Simonov — Head of the Center, Gazprom neft company group

* Автор, ответственный за переписку / Corresponding author