Прогноз обводненности на проектируемых к бурению скважинах методами машинного обучения

М.Р. Еникеев[1], М.Ф. Фазлытдинов[1], Л.В. Еникеева[2], И.М. Губайдуллин[2,3]

[1]Газпромнефть НТЦ, Набережная р. Мойки 75-79, лит. Д, Санкт-Петербург, Россия, 190000
[2]Институт нефтехимии и катализа - обособленное структурное подразделение Уфимского федерального исследовательского центра РАН, пр. Октября 141, Уфа, Россия, 450075
[3]Уфимский государственный нефтяной технический университет, Космонавтов 1, Уфа, Россия, 450062

Аннотация. За время эксплуатации нефтяных месторождений генерируется большое количество данных. Эти данные могут быть как интерпретированными специалистом, так и «сырыми», которые получены непосредственно с приборов, как структурированными, так и не структурированными, либо локально структурированными (то есть позволяют локально проводить анализ, но в данном виде не позволяют анализировать в совокупности с другими видами данных). Для получения из такого набора более информативных данных, которые позволяют принимать решения в процессе эксплуатации месторождения, требуется привлечение специалистов разных областей нефтяной отрасли. Поэтому возникает возможность и необходимость применения недетерминированных методов анализа полученных данных. В статье рассмотрено применение методов машинного обучения в задаче определения начальной обводненности по данным геофизических исследований скважин.

1. Введение

Начальная обводненность скважины – относительное содержание воды в добываемой жидкости, выраженное в процентах, при начале эксплуатации скважины. Она позволяет оценить целесообразность ввода в эксплуатацию нефтяной скважины. Одной из актуальных задач является прогноз обводненности по новым скважинам и выделение доли непроизводительной добычи/закачки при эксплуатации скважин, вскрывающих помимо целевого водонасыщенный горизонт. Динамика обводнения нефтяных скважин обусловливается характером обводнения нефтяных пластов. Характер обводнения пластов-коллекторов весьма различен и зависит от свойств продуктивных пластов, начальных условий залегания нефти в пласте. Также на характер заводнения и на динамику обводнения оказывает послойная и зональная неоднородность пластов. Интенсивность обводнения зависит от проницаемости пласта. Неравномерное обводнение пластов по их мощности и простиранию усиливается при высоком соотношении вязкости нефти и воды. Многие из этих факторов заложены в методах геофизического исследования скважин (ГИС).

ГИС — комплекс методов разведочной геофизики, используемых для изучения свойств горных пород в околоскважинном и межскважинном пространствах и для контроля технического

состояния скважин. Геофизические исследования скважин делятся на две группы — каротаж и скважинную геофизику.

Как правило, каротажные данные с месторождения представляют собой сильно и случайным образом флуктуирующие функции (рисунок 1).

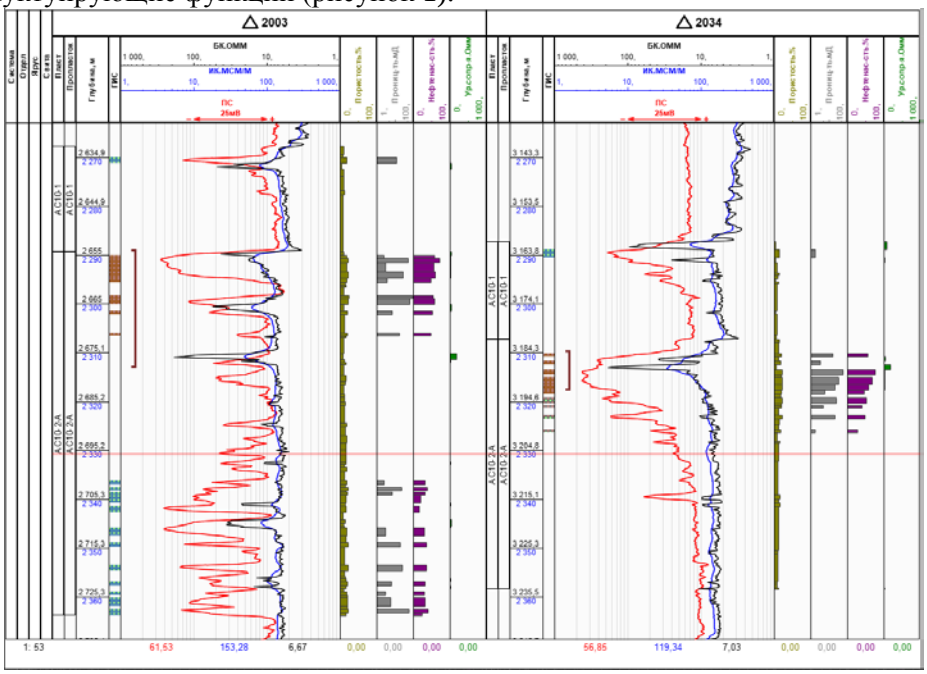


Рисунок 1. Пример каротажных данных: по оси y откладывается обычно глубина, по оси x — значения каротажа.

В работе [1] рассмотрена возможность применения аналого-статистических методов при прогнозе обводненности продукции добывающих скважин с учетом влияния геолого-технологических показателей. Одной из ключевых звеньев методики является построение для моделей-аналогов модифицированных зависимостей обводненности продукции скважин от степени выработки извлекаемых запасов нефти. Вывод об адекватности прогноза реальному геологическому объекту может быть сделан путем сопоставления фактической и прогнозной продуктивностей. Применение подобных методик позволяет существенно уменьшить ошибки прогноза дебитов и обводненности продукции новых скважин и увеличить надежность оперативного прогноза добычи. Дешененков [2] предложил способ прогноза продуктивности и начальной обводненности продукции нефтяных залежей с использованием ОФП по нефти и воде, промысловых геофизических данных и эффективной пористости. Для прогноза обводненности по данным ГИС в работе [3] используются формулы модели Кори, по которым определяются проницаемости для нефти и воды. Затем в перфорационных интервалах рассчитываются накопленные проницаемости по воде и нефти. После этого с учетом вязкостей обеих фаз оценивается коэффициент обводненности. В этих работах используются заранее заданные модельные формулы для сопоставления значений обводненности и каротажных данных, но не проведено глубокого анализа влияния различных типов каротажей на значения обводненности.

2. Постановка задачи

Как было упомянуто ранее, данные геофизических исследований являются основным источником информации о продуктивном пласте на стадии построения геологической модели и создании концепции разработки месторождения. Целью данной работы является разработка метода для прогноза обводненности по новым скважинам на основе каротажных данных.

Современное развитие методов машинного обучения позволяют эффективно решать широкий спектр задач в различных областях. Например, при прогнозировании возникновения инсульта

[4], прогнозирования состояния электромеханических систем прокатного производства [5], прогнозирование красных смещений галактик [6], а также в нефтяной отрасли, например, для интерпретации данных сейсмических исследований [7] и прогнозирования развития коррозии трубной стали [8]. Нефтяная отрасль является источником большого количества структурированных и неструктурированных данных. Разработаны и развиты большое количество инженерных(аналитических\эмпирических) методов изучения предметной области.
В терминах машинного обучения задача прогноза обводненности относится к задаче классификации — необходимо разделить множество объектов X (множество каротажных данных с месторождения) на M непересекающихся классов из Y (различные значения обводненности). Таким образом, определены понятия пространства объектов и пространства классов. Как известно, в задачах машинного обучения выделяют два этапа — этап обучения и этап применения. В данном случае обучающая выборка представляет собой набор проинтерпретированных геофизиком каротажных кривых, где для каждого элемента известно, к какому значению обводненности он относится. Неотъемлемым подготовительным этапом для работы алгоритма классификации является также отбор признаков объектов, о котором речь пойдет далее.

3. Реализация подхода

Задача сведена к признаковой задаче классификации. Сначала данные предобрабатывались: проводилась фильтрация по значениям (из данных удалялись значения nan, также были удалены заведомо неправильные значения (например, для aps значения > 1, для kint, значения < 0, и так далее)). Затем по каждому параметру строилось его признаковое описание. Признаки генерировались на основе разных подходов: аппроксимация данных, статистики, фурье-анализ (не вошли в финальное решение).
Далее каждый подход подробно описан и обоснован. Классификация выполнялась регрессионными алгоритмами, которые оценивали по данным каротажа значение обводненности. Были подробно исследованы возможности следующих обучающихся алгоритмов: случайный лес, бустинг над деревьями, нейросети.
Технически задача решалась следующим образом. С помощью библиотеки lasio и скрипта на python данные выгружались из las формата в csv (с данными значений параметров по глубине). Затем эти данные загружались в другом скрипте, производилась фильтрация данных, определялись значения кровли и подошвы, и генерировались признаки для задачи обучения, и набору признаков ласа ставились в соответствие значение обводненности (для генерации признаков методом Фурье использовался matlab). Данные разделялись для обучения и контроля, в соответствии 70/30. Данные признаки использовались для настройки регрессоров из библиотеки skit-learn и на нейросетях tensorflow и keras.

3.1. Загрузка и анализ данных

Для первоначального анализа данных, было решено проверить, нет ли явной зависимости значения обводненности (*wc*) от данных каротажа. Во время анализа данных, было замечено, что лучше анализировать область, лежащую между кровлей и подошвой. Для оценки было решено сравнить, масштабированные данные с усреднением значения в этой области для каждого параметра, для установления взаимосвязи между ними (рисунок 2). Из рисунка 2 можно сделать вывод, что выделение значимой информации не представляется возможным. Анализ зависимости в логарифмической шкале также не дал результатов. Поэтому было решено, проверить, насколько эффективны различные алгоритмы генерации признаков.

3.2. Генерация признаков

Задача классификация признаков сводилась к классической признаковой классификации, когда каждый объект (в данном случая кривая, либо набор кривых) описывался фиксированным набором вещественных признаков. Простейший пример признака – среднее значение на кривой. Однако, признаки должны быть выбраны так, чтобы максимально точно описать

сигнал, учесть физику описываемых процессов, также при генерации признаков могут быть использованы различные эвристики.
Для генерации признаков использовалось несколько подходов. Все они описаны ниже. Результатом такой генерации служит некий набор признаков, от десятков до сотен значений, в зависимости от метода.

3.2.1. Оценка признаков
Для оценки качества признаков использовался метод MAE (Mean absolute error – средняя абсолютная ошибка):

$$\text{MAE} = \frac{\sum_{i=0}^{N-1} |wc_{cor} - wc_{predict}|}{N}$$

где wc_{cor} – ожидаемое значение обводненности, $wc_{predict}$ – значение обводненности, предсказанное классификатором, N – количество скважин в контрольной выборке.

3.2.2. Признаки на основе кусочно-линейной интерполяции
Первым был опробован метод на основе кусочно-линейной интерполяции. Данные по выбранному параметру (кривой), после удаления некорректных значений, рассматривались в промежутке между кровлей и подошвой (в дальнейшем эту операцию будем обозначать - предобработкой), после чего генерировалось значение в k точках, с одинаковым шагом по глубине. Далее значения в этих k точках подаются в классификатор.

3.2.3. Статистические признаки
Сначала каждая кривая предобрабатывалась, затем вычислялись статистические признаки как для самого сигнала, так и для его производных. Генерация на этом этапе проводится в следующем порядке:

- Предобработка
- Для сигнала $(x_1, \dots, x_n)$, производной $(x_2 - x_1, x_n - x_{n-1})$ и модуля производной $(|x_2 - x_1|, |x_n - x_{n-1}|)$ вычисляются следующие значения признаков: среднее значение, стандартное отклонение, доля пересечений с уровнем *a* (*a = 0*, *a = mean*, *a = mean + std*)

Также был рассмотрен вариант с использованием признаков значений перцентилей (*p10, p50, p90* и т.д.), среднее значение и отклонение.

3.2.4. Выбор оптимальных признаков и параметра, по которому проводить обучение
Для выбора оптимального метода генерации признаком было решено провести классификацию ансамблем деревьев, с поиском оптимальных настроек для метода (изменение максимальной глубины дерева, количества деревьев в ансамбле решений, количество отбираемых признаков).
Для полного анализа оптимальной кривой ГИС для классификации, были проанализированы следующие кривые: 'kint', 'r05', 'r20', 'r14', 'r10', 'f07', 'f10', 'f14', 'r07', 'f20', 'f05', 'phit', 'mres', 'sg', 'kgl', 'sxwb', 'gz3', 'nphi', 'gz2', 'gz4', 'gz1', 'cild', 'prox', 'lld', 'gz7', 'aps', 'kps', 'gz5', так как эти кривые наиболее плотно наполнены данными.
Результаты сравнения работы алгоритмов генерации признаков для большинства кривых представлены на рисунке 2. В результате проведения данного теста, было решено использовать алгоритм, основанный на статистических признаках.
Среднее значения MAE по разным методам [0.127, 0.128, 0.130], и как мы видим, результат, практически не зависит от метода генерации параметров и от выбора кривой из данных каротажа.
Результат на контрольной выборке (30% от тестовой) по кривой gz5 представлен на рисунке 3. Признаки рассчитывались статистическим методом. Красным цветом изображены реальные значения обводненности, синим – предсказанные значения.
Аналогичная картина наблюдается и по другим каротажным кривым. Были проведены тесты по комбинации некоторых каротажных кривых при генерации параметров обучения. Существенного изменения результатов предсказания не наблюдалось.

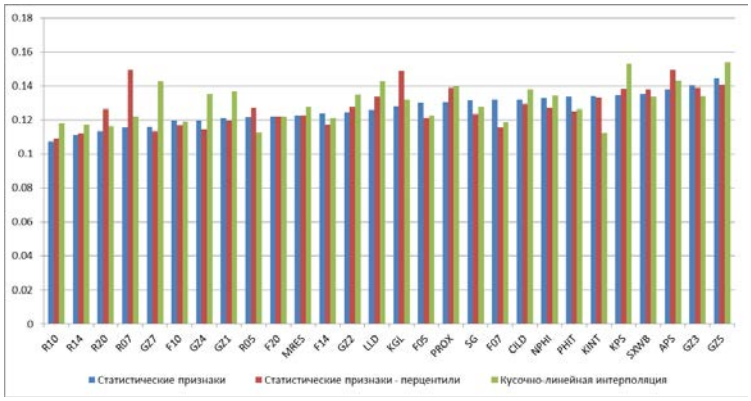


Рисунок 2. Сравнение методов генерации признаков для обучения.

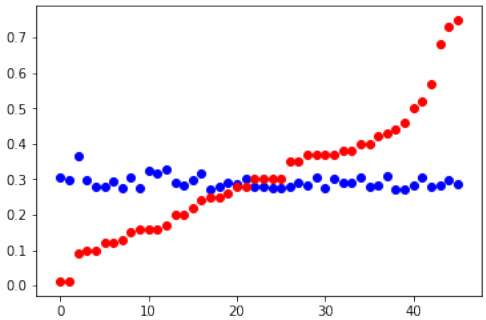
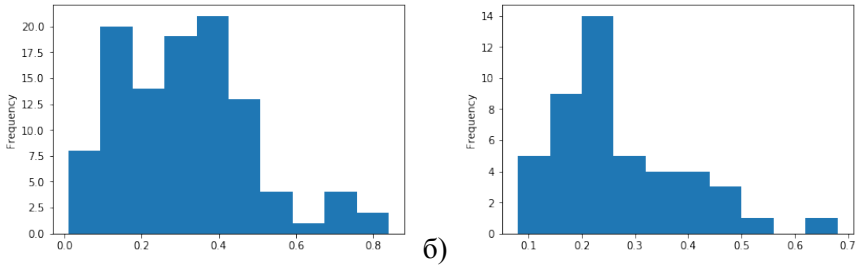


Рисунок 3. Результат работы классификатор по gz5 (красный цвет – реальные значения обводненности, синий - предсказанные).

Можно предположить, что полученный результат является следствием неравномерного распределения значений обводненности в обучающей выборке и классификатор стремится предсказать среднее значение обводненности во входных данных (рисунок 4).



а) б)

Рисунок 4. Гистограммы значений обводненности: для обучающей выборки (а) и для контрольной выборки (б).

Было решено проверить, как поведет себя классификатор на более обширной обучающей выборке.

4. Моделирование каротажных данных спектральным методом

Одним из основных требований к задаче для применения методов машинного является большой набор обучающей выборки. Возможные решения проблемы расширения обучающей выборки:

- Моделирование каротажных данных спектральным методом и интерполяция (аппроксимация) карты обводненности. В статьях [9 – 11] описан подход, который использовался для генерации каротажных данных месторождения.
- Использование каротажных данных и обводненности по нескольким «схожим» месторождениям (например, все месторождения Западной Сибири).

В данной работе было использовано первое решение.

Для генерации целевых значений обводненности необходима карта обводненности. Она была получена логарифмической аппроксимацией предоставленных данных начальной обводненности.

Результаты каротажных данных полученные с помощью спектрального моделирования представлены на рисунке 5 (синим цветом изображён исходный каротаж, красным – смоделированный) 105 скважина в моделировании не участвовала, и в итоге была корректно предсказана.

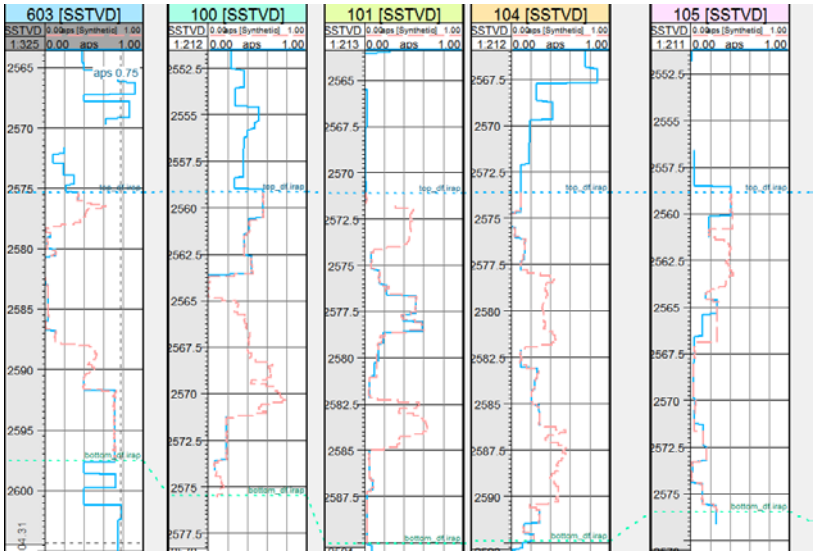


Рисунок 5. Результаты спектрального моделирования по кривой aps.

В результате моделирования обучающая выборка была расширена до 5349 скважин. То есть были получены расширенные данные для обучения классификатора.

4.1.1. Обучение на расширенной обучающей выборке

На данной выборке для генерации признаков использовалась кусочно-линейная интерполяция. В качестве исходных данных рассматривались кривые aps и kgl из РИГИС, так как в рамках поставленной задачи не было возможности проанализировать и расширить выборку по данным из ГИС.

В качестве инструмента классификации были рассмотрены ансамбль деревьев с поиском оптимальных настроек для метода и многослойная нейронная сеть.

Обученная модель проверялась на трех типах данных: контрольная выборка (выделенная из расширенного обучающего набора), значения *wc* на аппроксимированной карте (для исходных скважин) и реальные значения *wc* (для исходных скважин).

Ниже приведены результаты анализа работы классификаторов. Красным указаны – ожидаемые значения, синим – предсказанные значения. Для пояснения результата, к каждому графику дополнительно привязаны следующие значения: ‘MSE’, ‘MAE’, ‘R2 score’, ‘Explained variance score’:

- MSE – среднеквадратическая ошибка.

$$\mathrm{MSE} = \frac{\sum_{i=0}^{N-1}\left(wc_{cor} - wc_{predict}\right)^2}{N}$$

- MAE – средняя абсолютная ошибка, описана в 3.2.1.
- R2 score – коэффициент детерминации, является показателем качества регрессионной модели. Значение 1 соответствует идеальной прогнозирующей способности, а значение 0 соответствует константе модели, которая предсказывает среднее значение ответов.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=0}^{N-1}\left(wc_{cor} - wc_{predict}\right)^2}{\sum_{i=0}^{N-1}\left(wc_{cor} - wc_{mean}\right)^2}$$

- *Explained variance score* – объяснимая вариация.

$$\text{explained}_{\text{variance}} = 1 - \frac{Var\{wc_{cor} - wc_{predict}\}}{Var\{wc_{cor}\}}$$

В приведенных выше формулах: wc_{cor} – ожидаемое значение обводненности, $wc_{predict}$ –значение обводненности, предсказанное классификатором, N – количество скважин в контрольной выборке, $Var\{\}$ – дисперсия.

Результаты классификации на данных aps (ансамбль деревьев) представлены на рисунке 6.

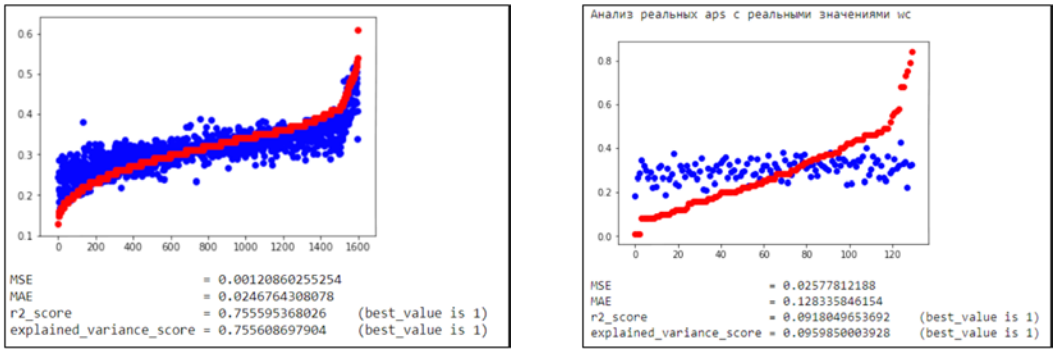


Рисунок 6. Результаты классификации ансамблем деревьев с генерацией признаков методом кусочной интерполяции по данным aps.

Результаты классификации на данных aps при помощи нейросети представлены на рисунке 7.

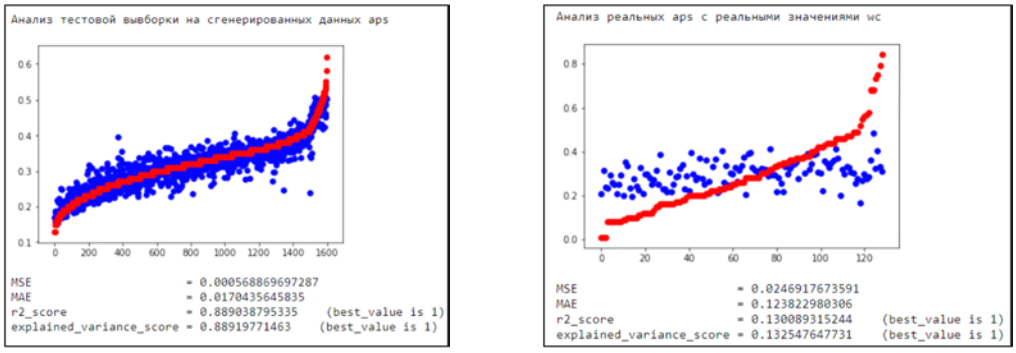


Рисунок 7. Результаты классификации нейросети с генерацией признаков методом кусочно-линейной интерполяции по данным aps.

Так как результаты работы классификаторов ансамблем деревьев и нейросети, отличаются незначительно, на примере aps, то дальнейшие проверки решено было проводить любым из классификаторов. Результаты классификации на данных kgl при помощи нейросети представлены на рисунке 8.

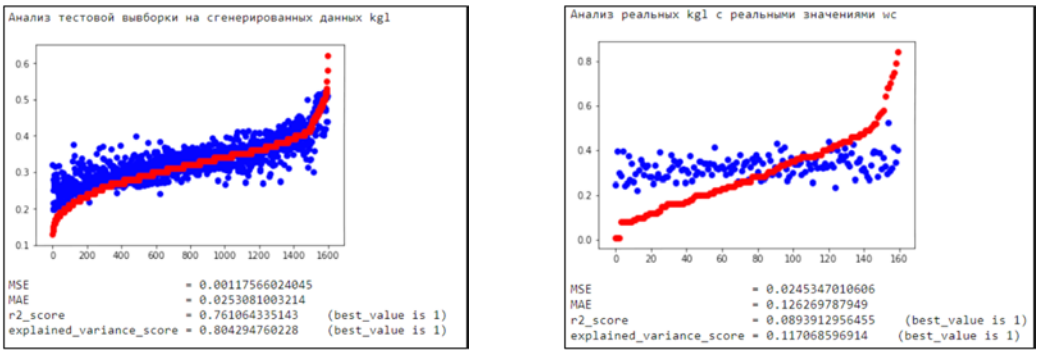


Рисунок 8. Результаты классификации нейросети с генерацией признаков методом кусочно-линейной интерполяции по данным kgl.

Были проанализированы результаты классификации на комбинации данных aps и kgl (рисунок 9). Как мы можем видеть, все модели показали на реальных данных результат одного порядка (на расширенных данных комбинация по kgl и aps показал лучшее предсказание, r2_score = 0.92 на контрольной выборке). Высокую точность на контрольной выборке, и гораздо худший прогноз на реальных скважинах. Например, r2_score = 0.92 и r2_score = -0.07 соответственно.

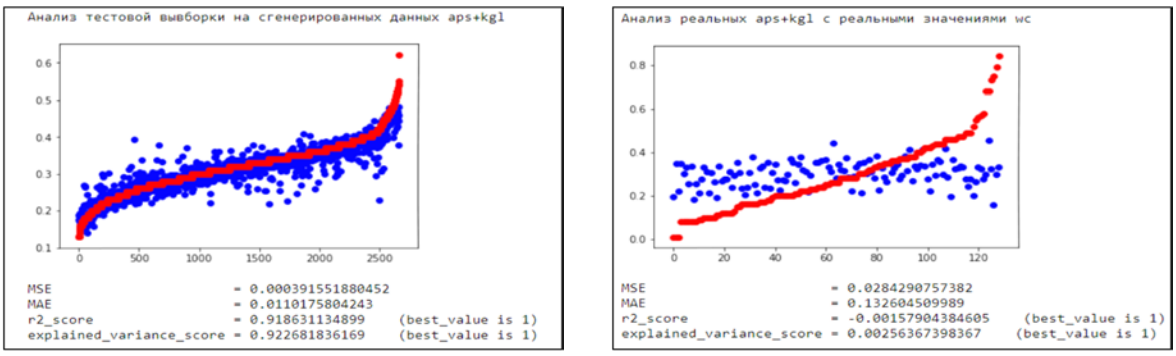
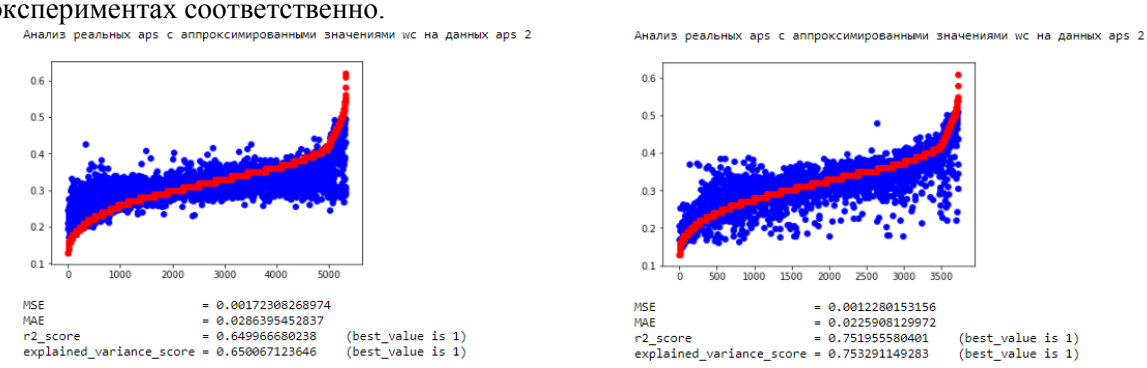


Рисунок 9. Результаты классификации нейросети с генерацией признаков методом кусочно-линейной интерполяции по объединенным данным kgl и aps.

Поэтому было решено проверить, как классификатор обученный на расширенных данных работает, для данных полученных в результате другого спектрального эксперимента, но с теми же начальными данными (рисунок 10). Из рисунка 12 и рисунков 7 – 8 следует, что на данных одного типа (смоделированные спектральным методом), классификатор показывает близкие результаты. Для нейросети r2_score = 0.89 и r2_score = 0.75, на первом и втором спектральном экспериментах соответственно.



а) б)

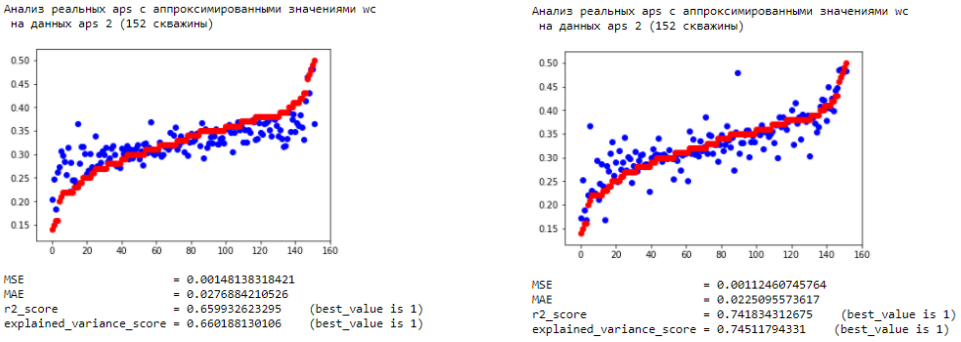
Рисунок 10. Результаты классификации (а) ансамблем деревьев, (б) нейросетью с генерацией признаков методом кусочно-линейной интерполяции по данным aps на части данных второго набора спектрального моделирования.

Точность предсказания r2_score обусловлены количеством или качеством данных, поэтому было решено проверить классификаторы на малом количестве тестовых данных, сгенерированных для эксперимента aps2. Были выбраны случайные 152 скважины. Как следует из рисунка 11 и рисунка 10, количество данных, не влияют на оценку точности прогноза (значения метрик точности совпадают). Возможно, что точки выбросов, лежат в малой окрестности реальных скважин, и если их не подавать на классификатор, то точность предсказания на расширенных данных будет точнее.

Причинами того, что классификатор показывает высокую точность на расширенных данных и низкую на реальных могут являться:

- Некорректный или не полный выбор кривых для генерации признаков.
- Предобработка данных практически не учитывала физические закономерности данных РИГИС.

- Способ генерации признаков. Возможно, выбранный нами способ, чувствителен к форме кривой, и поэтому выявляет закономерности генерации данных методом спектрального моделирования.



а) б)

Рисунок 11. Результаты классификации (а) ансамблем деревьев, (б) нейросетью с генерацией признаков методом кусочно-линейной интерполяции по данным aps на части данных второго набора спектрального моделирования (152 скважины).

5. Выводы

В работе рассмотрены основные аспекты обработки и анализа изображений в задаче исследования механизма коррозионных поражений:

- Проведен анализ и предобработка каротажных данных. Экспресс-анализ корреляции данных РИГИС и обводненности не выявил видимых закономерностей. Количество данных и распределение значений обводненности (см. рис. 4) является недостаточным для настройки регрессионного и мультиклассового классификатора.
- Исследованы 4 способа генерации признаков: кусочно-линейная интерполяция, 2 метода генерации статистических признаков и ДПФ-анализ. ДФП-анализ оказался плохо применим (возможно, авторы не смогли подобрать оптимальные коэффициенты, которые необходимо использовать для обучение). Метод основанный на кусочно-линейной интерполяции и методы на статистических признаках показали практически идентичные результаты.
- Рассмотрены 4 метода классификации: деревья, ансамбль деревьев, градиентный бустинг и многослойные нейросети. На сгенерированных признаках все методы показывали близкие результаты (кроме одиночного дерева), ансамбль деревьев и нейросети имели небольшое преимущество по времени обучения.
- Настроены классификаторы на данные РИГИС и ГИС и сделан вывод, что они стремятся предсказать среднее значение обводненности. Поэтому была расширена обучающая выборка на РИГИС, используя спектральное моделирование для данных каротажей и аппроксимацию обводненности, для увеличения разнообразия данных для обучения.
- Обучены классификаторы на расширенной выборке РИГИС, которые показали высокие точности предсказаний на контрольной расширенной выборке (один набора для обучения и один набор для установления точности, в каждом наборе около 5000, а обучение производилось на 3500). Однако точность этих классификаторов на исходных данных была низкой.

6. Литература

[1] Илюшин, П.Ю. Прогноз обводненности продукции добывающих скважин пермского края с применением аналого-статистических методов / П.Ю. Илюшин, С.В. Галкин / Вестник Пермского национального исследовательского политехнического университета. Геология, нефтегазовое и горное дело. – 2011. – Т. 10, № 1. – С. 76-84.

[2] Дешененков, И.С. Прогноз продуктивности и начальной обводненности нефтяных скважин одного из месторождений западной сибири по данным промысловой геофизики // Бурение и нефть. – 2013. – № 7-8. – С. 32-35.
[3] Алексеев, А.Д. Опыт и перспективы применения современных комплексов ГИС и ГДИС на месторождениях салымской группы / А.Д. Алексеев, А.А. Аниськин, Я.Е. Волокитин, М.С. Житный, Д.А. Карнаух, А.В. Хабаров // Инженерная практика. – 2011, № 11-12. – С. 62-75.
[4] Карп, В.П. Построение решающих правил в исследовании динамики космофизических показателей с целью прогнозирования ситуаций, провоцирующих возникновение эпизодов инсульта / В.П. Карп, Ю.А. Саяпина, Л.Г. Хетагурова, Н.К. Ботоева // Здоровье и образование в XXI веке. – 2012. – Т. 14, № 1. – С. 221-222.
[5] Кожевников, А.В. Применение методов машинного обучения в рамках прогнозирования состояния электромеханических систем прокатного производства / А.В. Кожевников, И.С. Илатовский, О.И. Соловьева / Вестник Череповецкого государственного университета. – 2017. – № 1. – С. 33-39.
[6] Герасимов, С.В. Применение платформы Microsoft Azure HDInsight для обработки и анализа больших массивов астрономических данных / С.В. Герасимов, А.В. Мещеряков / International Journal of Open Information Technologies. – 2017 – Т. 5, № 1. – С. 81-87.
[7] Краснов, Ф.В. Автоматизированное обнаружение геологических объектов в изображениях сейсмического поля с применением нейронных сетей глубокого обучения / Ф.В. Краснов, А.В. Буторин, А.Н. Ситников / Бизнес-информатика. – 2018. – № 2. – С. 7-16.
[8] Еникеев, М.Р. Информационно-вычислительная аналитическая система для оценки и прогнозирования коррозионных процессов на поверхности стали и алюминия / М.Р. Еникеев, И.М. Губайдуллин, М.А. Малеева / Системы и средства информатики. – 2017. – Т. 27, № 3. – С. 155-170.
[9] Байков, В.А. Новые подходы в теории геостатистического моделирования / В.А. Байков, Н.К. Бакиров, А.А. Яковлев // Вестник Уфимского государственного авиационного технического университета. – 2010. – Т. 14, № 2. – С. 209-215.
[10] Байков, В.А. Учет неоднородности при геолого-гидродинамическом моделировании Приобского месторождения / В.А. Байков, А.С. Бочков, А.А. Яковлев // Нефтяное хозяйство. – 2011. – № 5. – С. 50-54.
[11] Хасанов, М.М. Применение спектральной теории для анализа и моделирования фильтрационно-емкостных свойств пласта / М.М. Хасанов, Б.В. Белозеров, А.С. Бочков, О.С. Ушмаев, О.М. Фукс // Нефтяное хозяйство. – 2014. – № 12. – С. 60-64.

The apply of machine learning methods for watercut prediction on the projected wells

M.R. Enikeev[1], M.F. Fazlytdinov[1], L.V. Enikeeva[2], I.M. Gubaidullin[2,3]

[1]Gazpromneft STC, Moika River emb. 75-79, liter D, St Petersburg, Russia, 190000
[2]Ufa State Petroleum Technological University, Kosmonavtov St. 1, Ufa, Russia, 450062
[3]Institute Petrochemistry and Catalysis - Subdivision of the Ufa Federal Research Centre of RAS, pr. Oktyabria 141, Ufa, Russia, 450075

Abstract. During operation of oil fields a large number of data is generated. These data can be as interpreted by the expert or received directly from devices, both structured, and not structured, or locally structured. Receiving from such data set of more informative data which allow to make decisions in use of the field requires involvement of experts of different areas of the oil industry. Therefore there is an opportunity and need of application of nondeterministic methods of the analysis of the obtained data. For example, such task is determination of initial water content for the designed well. In article application of methods of machine learning in a problem of determination of initial water content according to geophysical surveys of wells is considered.