Министерство образования и науки Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт
Высшая школа теоретической механики и математической физики

Работа допущена к защите
Директор ВШТМиМФ,
д.ф.-м.н., чл.-корр. РАН
_____ А.М. Кривцов
«___»_____2022 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

ПРОГНОЗИРОВАНИЕ ОБВОДНЕННОСТИ СКВАЖИН ПО ГЕОЛОГО-ФИЗИЧЕСКИМ ХАРАКТЕРИСТИКАМ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

по направлению подготовки
01.03.03 Механика и математическое моделирование
профиль
01.03.03_03 «Математическое моделирование процессов нефтегазодобычи»

Выполнил
студент гр.5030103/80301 Д.Э. Давлетшин

Руководитель
Доцент ВШТМиМФ, к.ф.-м.н. И.В. Курта

Консультант
ассистент ВШТМиМФ А.Д. Ершов

Санкт-Петербург
2022

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт

Высшая школа теоретической механики

УТВЕРЖДАЮ

Директор высшей школы

теоретической механики и

математической физики

_____ А.М. Кривцов

« » _____20__ г.

ЗАДАНИЕ

по выполнению выпускной квалификационной работы

студенту Давлетшину Дмитрию Эдуардовичу, гр. 5030103/80301

1. Тема работы: ПРОГНОЗИРОВАНИЕ ОБВОДНЕННОСТИ СКВАЖИН ПО ГЕОЛОГО-ФИЗИЧЕСКИМ ХАРАКТЕРИСТИКАМ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ
2. Срок сдачи студентом законченной работы:

3. Исходные данные по работе:

4. Содержание работы (перечень подлежащих разработке вопросов):

5. Перечень графического материала (с указанием обязательных чертежей): не предусмотрены

6. Консультанты по работе: ассистент ВШТМ А.Д. Ершов

7. Дата выдачи задания 01.02.2022

Руководитель ВКР И.В. Курта

(подпись) инициалы, фамилия

Задание принял к исполнению 01.02.2022

(дата)

Студент Д.Э. Давлетшин

(подпись) инициалы, фамилия

РЕФЕРАТ

На 50 с., 30 рисунков, 7 таблиц.

АППРОКСИМАЦИЯ ДАННЫХ, КЛАСТЕРИЗАЦИЯ, МАШИННОЕ ОБУЧЕНИЕ, МУЛЬТИКЛАССОВАЯ КЛАССИФИКАЦИЯ, К-СРЕДНИХ, АГЛОМЕРАТИВНЫЙ ИЕРАРХИЧЕСКИЙ МЕТОД, ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, СЛУЧАЙНЫЙ ЛЕС, ЭКСТРЕМАЛЬНЫЕ СЛУЧАЙНЫЕ ДЕРЕВЬЯ, ГРАДИЕНТНЫЙ БУСТИНГ

В данной работе было проведено прогнозирование обводненности скважин с использованием различных методов аппроксимации, кластеризации и машинного обучения. Работа состоит из предобработки данных, разработки алгоритмов, разработки метрик для их оценки, валидации и сравнения моделей.

THE ABSTRACT

50 pages, 30 pictures, 7 tables.

DATA APPROXIMATION, CLUSTERING, MACHINE LEARNING, MULTICLASS CLASSIFICATION, K-MEANS, AGGLOMERATIVE HIERARCHICAL CLUSTERING, LOGISTIC REGRESSION, RANDOM FOREST, EXTRA-TREES, GRADIENT BOOSTING

In this work, the prediction of the water cut levels of wells was carried out using various approximation, clustering, and machine learning technics. The work consists of data preprocessing, development of algorithms and metrics for evaluation, validation and model comparison.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ **6**

ГЛАВА 1. ОБЩИЕ СВЕДЕНИЯ **8**

1.1. Виды скважин 8

1.2. Геолого-физические характеристики скважин 9

1.3. Машинное обучение 11

ГЛАВА 2. АППРОКСИМАЦИЯ И КЛАСТЕРИЗАЦИЯ **13**

2.1. Входные данные 13

2.2. Очистка данных и фильтрация скважин 14

2.3. Задача аппроксимации **16**
2.3.1. Обводненность по теории Баклея-Леверетта 18
2.3.2. Экспоненциальное и логистическое распределение 18
2.3.3. Распределения Гомпертца и Вейбулла 19
2.3.4. Распределение Ричардса 21

2.4. Нелинейный метод наименьших квадратов 21

2.5. Оценка качества аппроксимации 22

2.6. Задача кластеризации **25**
2.6.1. Метод k-средних 27
2.6.2. Агломеративный иерархический метод (АИМ) 28

2.7. Оценка качества кластеризации 30

ГЛАВА 3. ПРОГНОЗИРОВАНИЕ И ВАЛИДАЦИЯ МОДЕЛЕЙ **36**

3.1. Входные данные по геолого-физическим характеристикам 36

3.2. Обработка признаков 36

3.3. Задача прогнозирования **38**
3.3.1. Логистическая регрессия (One-vs-Rest) 40
3.3.2. Random forest 41
3.3.3. Extra-Trees 42
3.3.4. Градиентный бустинг 43

3.4. Метрики качества 43

3.5. Сравнение результатов машинного обучения 45

ЗАКЛЮЧЕНИЕ 46

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 48

ВВЕДЕНИЕ

Актуальность

Важной характеристикой при добыче нефти из скважин является показатель обводненности. Под обводненностью скважины подразумевают содержание воды в продукции скважины, которое определяется как отношение дебита воды к сумме дебитов нефти и воды. Основной период эксплуатации скважин отличается довольно постоянной и низкой обводненностью, а по мере разработки залежи обводненность скважин возрастает.

Производительность и обводненность скважин сильно коррелируют с величиной затрат на добычу. Чем больше временной промежуток, на котором сохраняются низкие уровни обводненности, тем большую прибыль компания может получить.

Прогнозирование обводненности играет большую роль в оценке рентабельности нефтяной скважины. Оно позволяет еще до начала эксплуатационных работ, имея лишь геолого-физические характеристики самой скважины, оценить ее обводненность. Вследствие чего нефтедобывающая компания может заранее предпринять верные решения по разработке месторождения, чтобы наиболее оптимально и эффективно задействовать свои ресурсы для добычи нефти.

Цель работы

Целью данной работы является построение модели машинного обучения, позволяющей спрогнозировать обводненность нефтяной скважины по ее геолого-физическим характеристикам на основе реальных исторических данных по множеству скважин, расположенных на территории России.

Задачи работы

1. Предобработка и фильтрация данных;
2. Аппроксимация обводненности скважин;
3. Кластеризация полученных аппроксимаций;
4. Обучение и валидация моделей на размеченных данных;

5. Сравнение моделей;
6. Анализ результатов.

ГЛАВА 1. ОБЩИЕ СВЕДЕНИЯ

1.1. Виды скважин

Проектирование скважины – один из основных подготовительных этапов при разработке месторождения нефти или газа, от него напрямую зависят объемные показатели добываемых ресурсов при бурении скважины. К наиболее распространенным относятся скважины следующих типов [16] (рис.1):

- наклонно-направленные;
- горизонтальные;
- вертикальные;
- многоствольные.

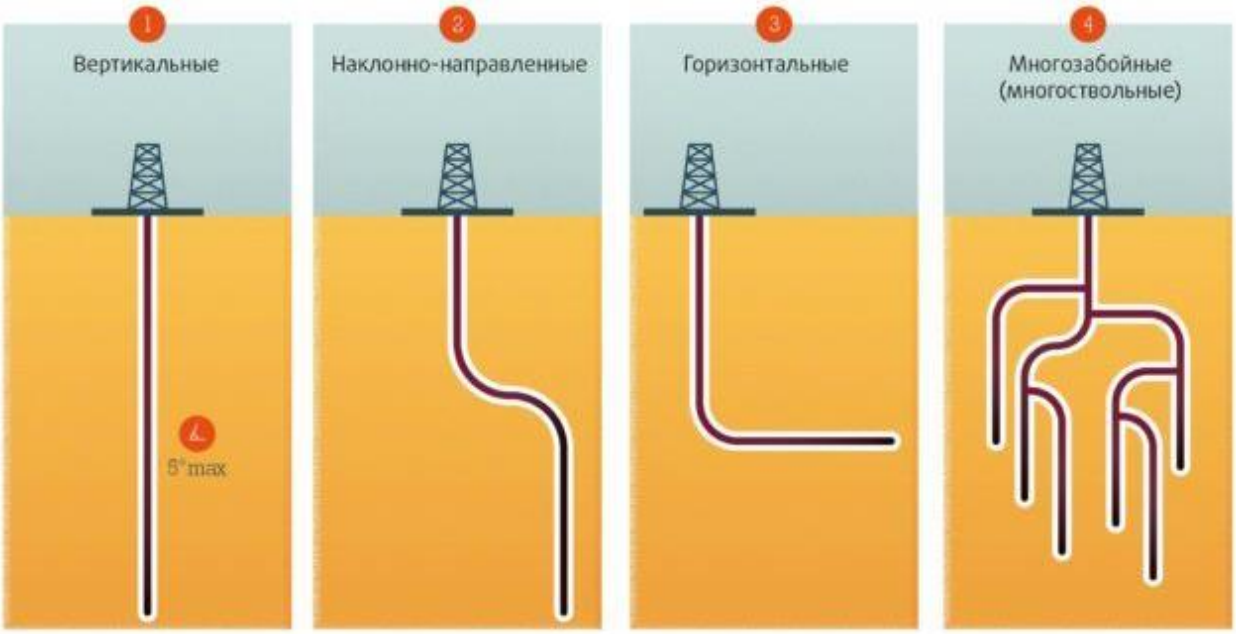

Рисунок 1. Типы скважин в зависимости от условий месторождения.

Наклонными считаются скважины, отклонение которых от вертикали составляет [17]:

- более 2° при колонковом бурении;
- более 6° при глубоком бурении скважин.

Наклонно-направленными называются скважины, направление которых во время процесса бурения строго контролируется. Соответственно, горизонтально направленное бурение является частным случаем наклонного,

т.е. бурением под углом в 90°. Также существуют многоствольные скважины, состоящие из основного ствола, из которого пробурен 1 или более боковых стволов.

Направленное и горизонтальное бурение — это методы, используемые для увеличения добычи и достижения показателей, которые не могут быть достигнуты при вертикальном бурении. Тем не менее большинство скважин, пробуренных для добычи нефти, природного газа или других подземных целей, представляют собой вертикальные скважины, пробуренные прямо в землю. Когда направленное бурение сочетается с гидроразрывом пласта, некоторые горные породы, которые были неэффективными при вертикальном бурении, могут стать гораздо более продуктивными в добыче нефти или природного газа. Наклонно-направленное и горизонтальное бурение используются для достижения целей под пролегающими пластами, уменьшения площади разработки газовых месторождений, увеличения длины продуктивной зоны в скважине, преднамеренного пересечения трещин, строительства разгрузочных скважин и прокладки инженерных коммуникаций под землями, где раскопки невозможны или чрезвычайно дороги.

1.2. Геолого-физические характеристики скважин

Не менее важной характеристикой при оценке продуктивности скважин являются их геолого-физические характеристики. В данной работе будут использоваться лишь некоторые из них.

Коллекторы – это горные породы, обладающие способностью вмещать нефть, газ и воду, и отдавать их при разработке. Первая используемая характеристика – это пористость. Пористость является одним из свойств коллекторов и отражает отношение объема пор к общему объему горной породы, выражается в процентах [18].

Мощность горных пород – толщина пласта или комплекса геологических отложений, замеренная по кратчайшему расстоянию (нормали) между его кровлей и подошвой [20]. Вторая характеристика – эффективная

мощность пласта, под которой подразумевают суммарную мощность прослоев коллекторов в пласте, выражается в метрах.

Третья характеристика – это проводимость пласта. Проводимость T является фильтрационным параметром пласта и выражается как расход жидкости через единицу поперечного сечения водоносного пласта мощностью m при единичном напорном градиенте [13].

$$T = km, \qquad [T] = \frac{\text{м}^2}{\text{сут}}$$

где k – это коэффициент фильтрации слоя

Для многослойного пласта, состоящего из n слоев, проводимость выглядит следующим образом:

$$T = \sum_{i=1}^{n} k_i m_i$$

где k_i – это коэффициент фильтрации i-го слоя,

m_i – мощность i-го слоя.

Последняя используемая характеристика – это проницаемость горных пород. Под проницаемостью пласта подразумевают способность пород к пропусканию жидкости и газа при перепаде давления. При относительно небольших перепадах давления в нефтяных пластах многие породы в результате незначительных размеров пор оказываются практически непроницаемыми для жидкостей и газов (глины, сланцы и т.д.) [19]. Для оценки проницаемости горных пород применяют открытый в 1856 г линейный закон фильтрации Дарси, который установил зависимость скорости фильтрации жидкости от градиента давления:

$$k = \mu \frac{q_\text{ф} L}{\Delta P F}$$

где k– это проницаемость пористой среды [м^2], μ – динамическая вязкость флюида [Па · с], $q_\text{ф}$ – объемный расход флюида (дебит) $\left[\frac{\text{м}^3}{\text{с}}\right]$, L –

длина образца пористой среды [м], ΔP – перепад давления [Па], F – площадь фильтрации [м^2].

1.3. Машинное обучение

Машинное обучение – это раздел искусственного интеллекта (ИИ) и информатики, который фокусируется на использовании данных и математических алгоритмов для имитации того как обучается человек, постепенно повышая свою точность. При помощи статистических методов алгоритмы тренируются в задачах классификации или прогнозирования, раскрывая ключевые особенности входных данных в рамках проектов автоматизированного анализа больших данных. Эти закономерности впоследствии влияют на принятие решений в приложениях и компаниях, в идеале влияя на ключевые показатели роста [14]. Основные задачи делятся на 2 типа:

- обучение с учителем – основано на выявлении закономерностей в данных, когда для множества объектов (признаков) имеется множество возможных ответов (откликов);
- обучение без учителя – методы обучения, в которых известны только описания множества объектов, и требуется обнаружить внутренние взаимосвязи.

В случае обучения с учителем задается множество объектов X, множество Y допустимых ответов, и существует целевая функция $f: X \to Y$, ее значения $y_i = f(x_i)$ известны только на ограниченном множестве $x_1, \dots, x_l \subset X$ и $X^l = (x_i, y_i)_{i=1}^l$.

Обучение с учителем также разделяется на несколько типов:

- классификация – множество возможных ответов конечно $Y = \{1, \dots, N\}$;
- регрессия – возможные ответы являются множеством действительных чисел $Y = \mathbb{R}$

Таким образом, в данной работе речь пойдет именно о горизонтальных и наклонно-направленных скважинах, обводненность которых и будет прогнозироваться на основе вышеуказанных геолого-физических характеристик. В качестве моделей машинного обучения будут использоваться как классические модели классификации, так и их модификации с предварительной разметкой данных при помощи аппроксимирующих распределений и моделей обучения без учителя. Соответственно, результаты прогнозирования по обоим типам скважин будут рассматриваться в отдельности друг от друга.

ГЛАВА 2. АППРОКСИМАЦИЯ И КЛАСТЕРИЗАЦИЯ.

2.1. Входные данные

На рисунке 2 представлен вид основной таблицы с данными по обводненности скважин, которую мы получаем на вход. Представленные в ней колонки несут следующий смысл:

- скважина – идентификационный номер скважины;
- дата – год и месяц, на которые был сделан замер;
- нефть – объем добытой нефти в дебете за данный период;
- вода – объем воды в дебете за данный период;
- Qcum – кумулятивная жидкость, рассчитываемая как сумма нефти и воды;
- Qwatercut – отношение дебета воды к сумме нефти и воды.

Общее количество наклонно-направленных скважин: 247

Общее количество горизонтальных скважин: 99

Скважина	Дата	Нефть	Вода	Жидкость	Qcum	Qwatercut
2692	2007-10-01	383	286.012	669.012000	669.012	0.427514
2692	2007-11-01	496	492.69	988.690000	1657.702	0.498326
2692	2007-12-01	469	608.485	1077.485000	2735.187	0.564727
2692	2008-01-01	610	949.421	1559.421000	4294.608	0.608829
2692	2008-02-01	676	936.926	1612.926000	5907.534	0.580886
...	...	...	...	...	...	...
3410	2021-02-01	190.074213	735.074573	925.148786	103806.829176	0.794547
3410	2021-03-01	262.702008	1033.705734	1296.407743	105103.236918	0.797362
3410	2021-04-01	111.130525	1038.666249	1149.796774	106253.033692	0.903348
3410	2021-05-01	112.691126	855.068804	967.759930	107220.793622	0.883555
3410	2021-06-01	44.120119	283.37772	327.497839	107548.291461	0.865281

Рисунок 2. Вид входных данных по обводненности скважин.

2.2. Очистка данных и фильтрация скважин

Реальные данные собираются для последующей обработки из различных источников и процессов. Они могут содержать ошибки, повреждения, дубликаты, что негативно влияет на качество набора данных. К типичным проблемам с качеством данных относятся:

- неполнота: данные не содержат атрибутов, или в них пропущены значения;
- шум: данные содержат ошибочные записи или выбросы;
- несогласованность: данные содержат конфликтующие между собой записи или расхождения.

Качественные данные – это необходимое условие для дальнейшего их применения и создания на их основе качественных моделей прогнозирования. Введем правила, по которым будем улучшать выборку исходных данных:

1. отфильтруем скважины, по которым имеется меньше 6 записей, т.е. период работы которых составил 6 месяцев;
2. отфильтруем скважины, дающие 0 при суммировании по Qcum, т.е. таких, у которых были большие простои;
3. отфильтруем скважины, у которых меньше 6 уникальных значений Qcum, в таких случаях чаще всего имеем дело с шумными данными.

Для визуализации результата предобработки выведем графики обводненности до (рисунок 3) и после (рисунок 4) по 16 случайным скважинам.

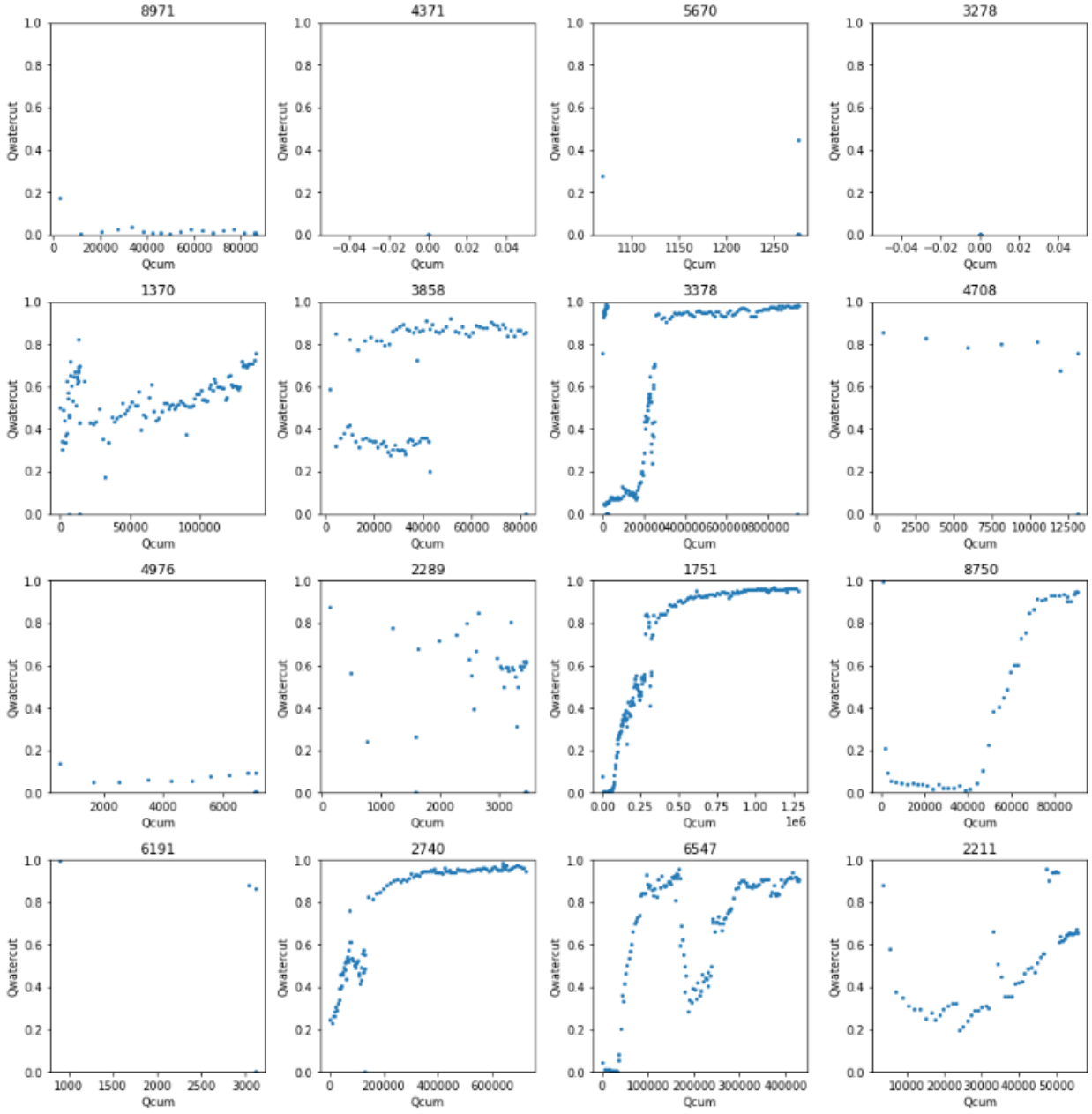

Рисунок 3. Графики обводненности скважин до предобработки.

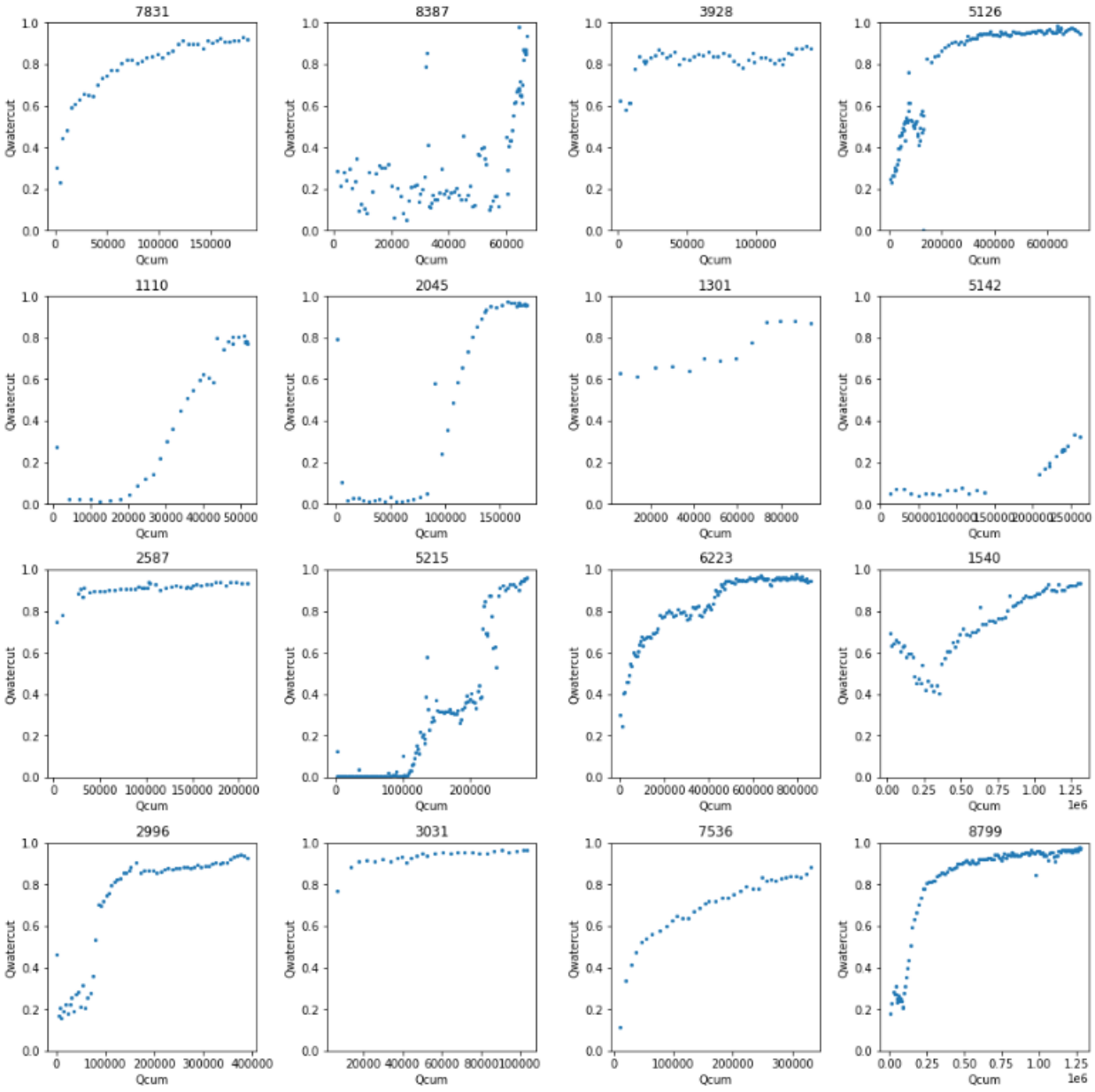

Рисунок 4. Графики обводненности скважин после предобработки.

Видим, что удалось избавиться от большинства часто встречающихся ошибок в данных. Остальные скважины отфильтруем вручную, если они не подходят по графикам обводненности.

2.3. Задача аппроксимации

Задача аппроксимации заключается в том, чтобы заменить сложные объекты близкими к ним, но более простыми. Наиболее часто аппроксимация применяется в случаях, когда исходную модель сложно использовать или наблюдаются проблемы с качеством данных. Таким образом, данный подход позволяет свести к минимуму влияние выбросов в

данных, а также обобщить их, что в целом важно для последующего применения машинного обучения.

Тип используемой аппроксимации зависит от исходных данных, необходимого качества приближения, влияния входных данных на искомое решение задачи и временных затрат как на расчеты, так и на усилия со стороны исследователя.

В данной работе будут аппроксимироваться данные, представляющие собой график обводненности, т.е. отношение объемной доли потока вытесняемой жидкости (воды) к суммарному потоку двух фаз (нефти и воды). Пример графика обводненности скважины и соответствующей ему аппроксимации на основе исходных данных показан на рисунке 5. Для более точной аппроксимации будем использовать известные распределения в форме кумулятивных функций, которые позволят учесть индивидуальные особенности распределений каждой из скважин засчет вариации входящих в уравнения параметров.

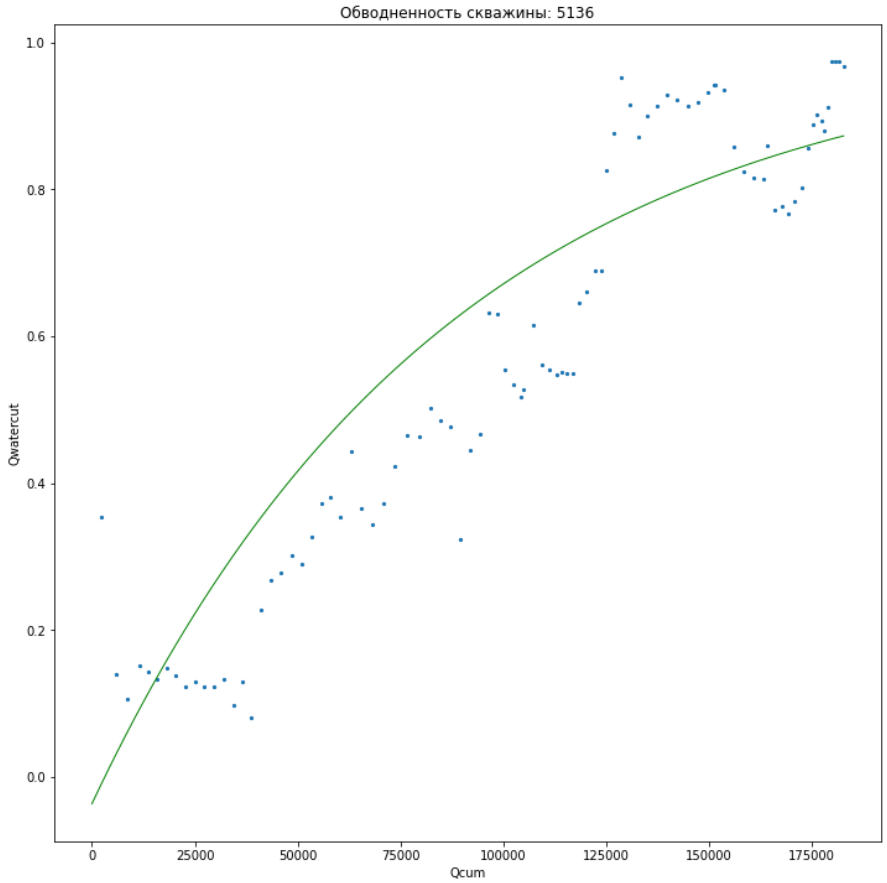

Рисунок 5. График обводненности скважины и соответствующей ему аппроксимации.

2.3.1. Обводненность по теории Баклея-Леверетта

В случае одномерного течения несжимаемых несмешивающихся жидкостей в классической теории вытеснения наибольшее распространение получила теория Баклея-Леверетта. Предложенная функция, описывающая постепенное обводнение скважины выглядит следующим образом [2]:

$$f(\sigma) = \frac{k_{\text{в}}(\sigma)}{k_{\text{в}}(\sigma) + \frac{\mu_{\text{в}}(\sigma)}{\mu_{\text{н}}(\sigma)} k_{\text{н}}(\sigma)}$$

где $k_{\text{в}}, k_{\text{н}}$ – относительные фазовые проницаемости воды и нефти;

σ – насыщенность.

Физический смысл этой функции заключается в том, что она представляет собой отношение скорости фильтрации вытесняющей фазы к суммарной скорости, и равна объемной доле потока вытесняющей жидкости (воды) в суммарном потоке двух фаз. Таким образом, функция Баклея-Леверетта определяет полноту вытеснения и характер распределения газоконденсатонасыщенности по пласту [21]. На рисунке 6 показаны функции Баклея-Леверетта и ее производной.

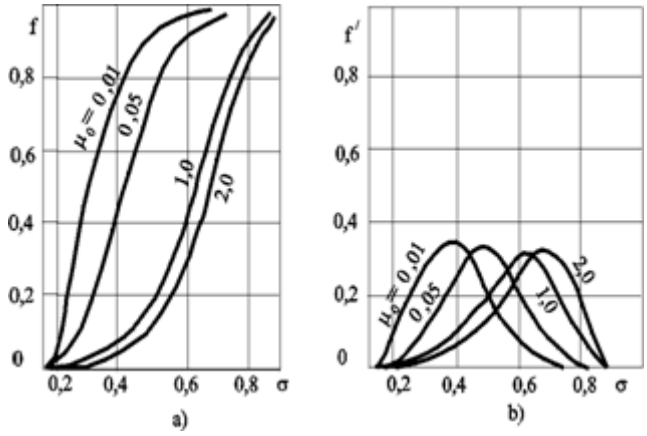

Рисунок 6. Графики функции Баклея-Леверетта (a) и ее производной (b) для различных отношений вязкости.

2.3.2. Экспоненциальное и логистическое распределение

Экспоненциальное распределение можно описать следующей функцией:

$$f(t) = 1 - (1 - \beta)e^{-ct}$$

где $\beta > 0$ – параметр масштаба; $c > 0$ – параметр скорости распределения.

Логистическое распределение – частный случай экспоненциального распределения и имеет следующее представление:

$$f(t) = \frac{1}{1 + \beta e^{ct}}$$

где $\beta > 0$ – параметр масштаба; $c > 0$ – параметр скорости распределения.

Приведем пример применения данных распределений. С помощью обоих распределений можно смоделировать рост населения какой-либо страны, разница будет заключаться в том, что логистическое распределение будет уже учитывать, например, ограниченность территории и ресурсов страны, что ограничит бесконечный рост населения.

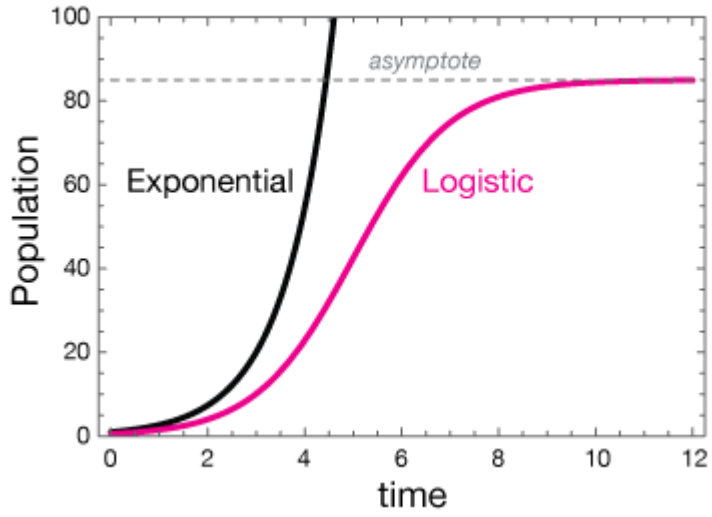

Рисунок 7. Пример экспоненциального и логистического распределений.

2.3.3. Распределения Гомпертца и Вейбулла

Данные распределения схожи с вышеупомянутыми распределениями. Распределение Гомпертца имеет следующий вид [3]:

$$f(t) = 1 - e^{(-\eta(e^{\beta t} - 1))}$$

где $\beta > 0$ – параметр масштаба; $\eta > 0$ – параметр формы.

Распределение Вейбулла [11]:

$$f(t) = 1 - e^{(-\beta t)^{\eta}}$$

где $\beta > 0$ – параметр масштаба; $\eta > 0$ – параметр формы.

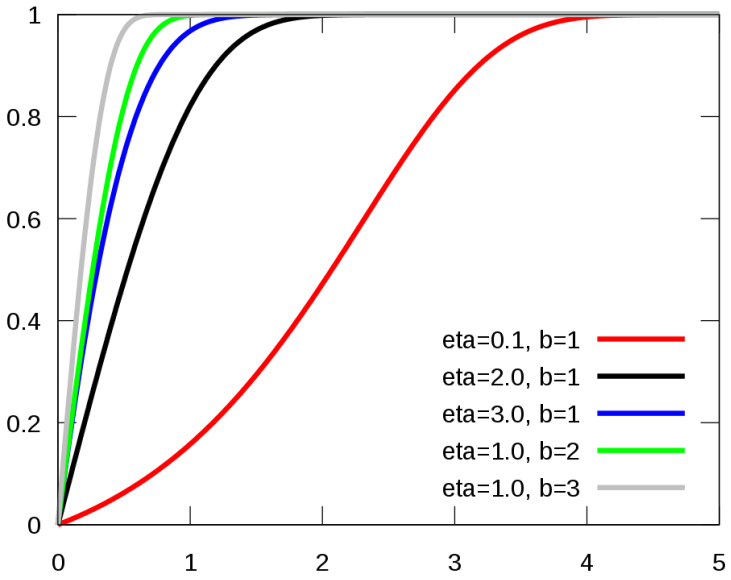

Рисунок 8. Распределение Гомпертца при различных параметрах.

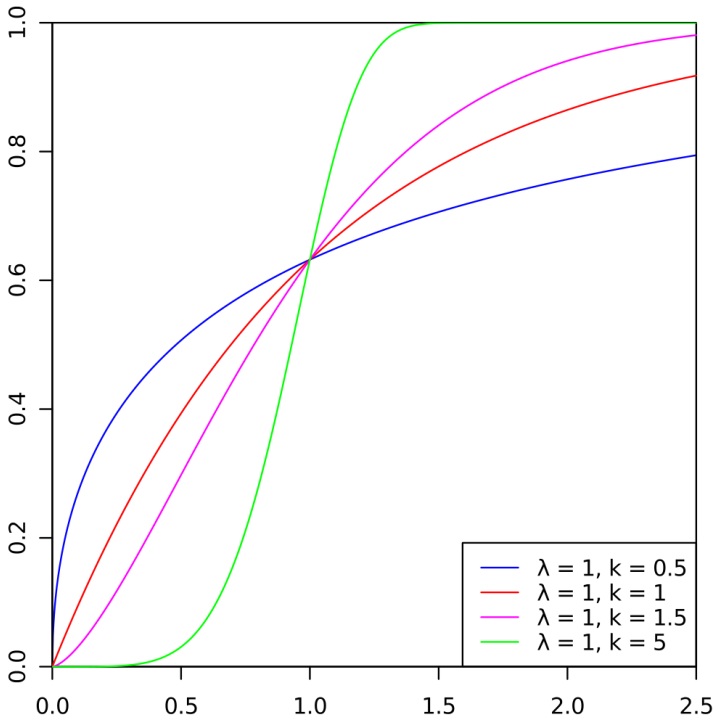

Рисунок 9. Распределение Вейбулла при различных параметрах.

С точки зрения статистики наибольший интерес к подобным функциям возникает при оценке продолжительности работоспособности различных устройств. В таком контексте экспоненциальное распределение имеет

постоянную частоту поломок, распределение Гомпертца – возрастающую, а Вейбулла – уменьшающуюся [12].

2.3.4. Распределение Ричардса

Обобщенная логистическая функция, также известная как распределение Ричардса, выглядит следующим образом [9]:

$$f(t) = A + \frac{K - A}{(1 + Qe^{-Bt})^{\frac{1}{\nu}}}$$

где A – нижняя асимптота; K – верхняя асимптота;

B – темп роста; $\nu > 0$ – отвечает за то у какой асимптоты будет наблюдаться наибольший рост функции; Q – значение в $f(0)$.

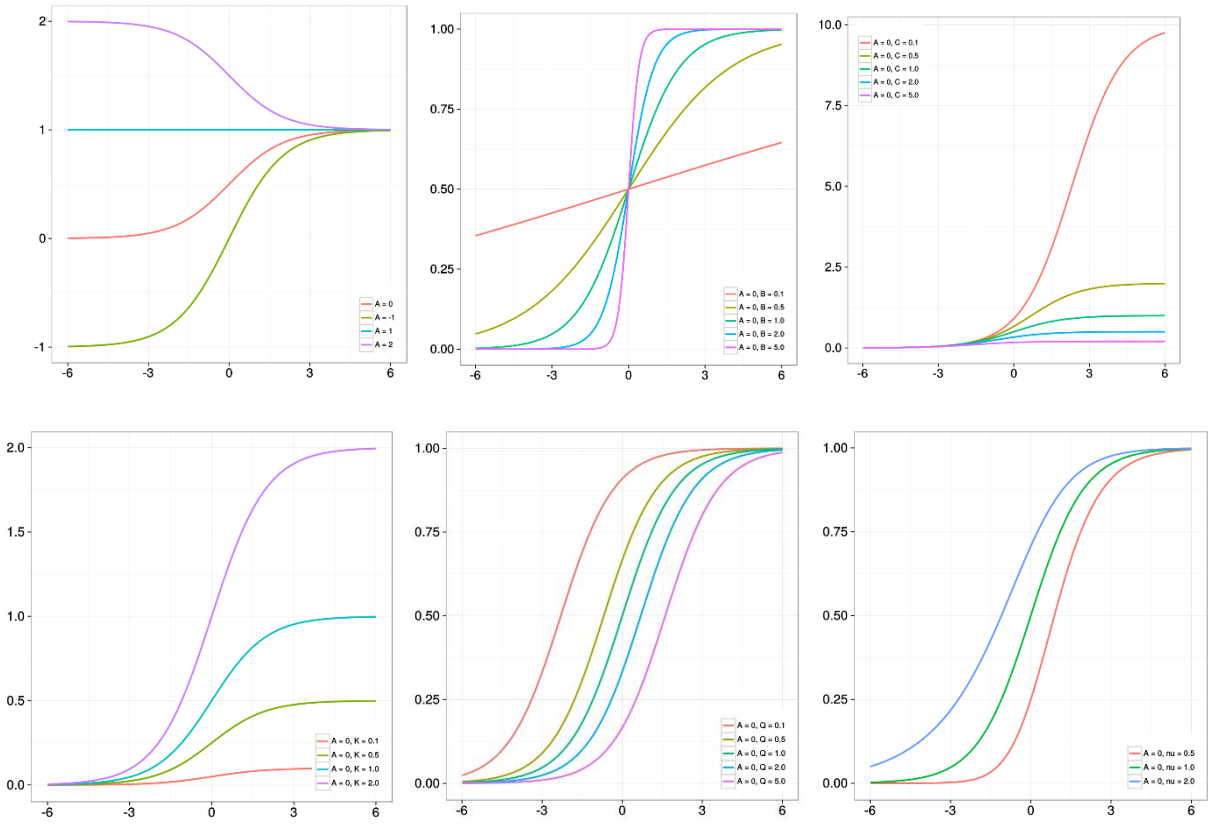

Рисунок 10. Распределение Ричардса при варьировании различных параметров.

2.4. Нелинейный метод наименьших квадратов

Наиболее точная аппроксимация распределения обводненности скважины с помощью какого-либо распределения из пункта 2.3 достигается при определенной комбинации параметров, входящих в это распределение.

Для нахождения такой комбинации будем использовать нелинейный метод наименьших квадратов (НМНК). Опишем логику работы НМНК [10].

Допустим у нас есть m точек $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ и некоторая кривая, зависящая от n параметров β вида $y = f(\beta_1, \dots \beta_n; x)$, причем $m \geq n$. Требуется найти вектор $\beta = [\beta_1, \dots \beta_n]$, минимизирующий квадратичную ошибку:

$$S = \sum_{i-1}^{m} r_i^2, \qquad r_i = y_i - f(\beta_1, \dots \beta_n; x)$$

Минимум S достигается при обращении градиента в ноль:

$$\frac{\partial S}{\partial \beta_i} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0, \qquad j = 1, \dots, n$$

2.5. Оценка качества аппроксимации

Таким образом, для каждой аппроксимирующей функции подберем оптимальную комбинацию параметров по НМНК. Затем из всех этих аппроксимаций необходимо выбрать такую, которая наиболее точно описывает распределение обводненности скважины.

Существует множество метрик, по которым можно выбрать наиболее точное аппроксимирующее распределение, таких как средняя квадратичная ошибка (MSE):

$$MSE = \frac{1}{k} \sum_{i=1}^{k} (y_i - y_i')^2$$

где $k -$ количество сравниваемых точек.

Или корень из средней квадратичной ошибки (RMSE):

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (y_i - y_i')^2}$$

где $k -$ количество сравниваемых точек.

Но в данном случае наиболее подходящей будет метрика средней абсолютной ошибки (MAE), т. к. она наиболее устойчива к выбросам в данных [22]:

$$MAE = \frac{1}{k}\sum_{i=1}^{k}|y_i - y_i'|$$

где k – количество сравниваемых точек.

На рисунке 11 представлены результаты аппроксимаций. Красная кривая аппроксимации обозначает, что скважина не подходит по заданному критерию не убывания аппроксимации при возрастании объема кумулятивной жидкости, зеленая – скважина подходит и будет включена в выборку.

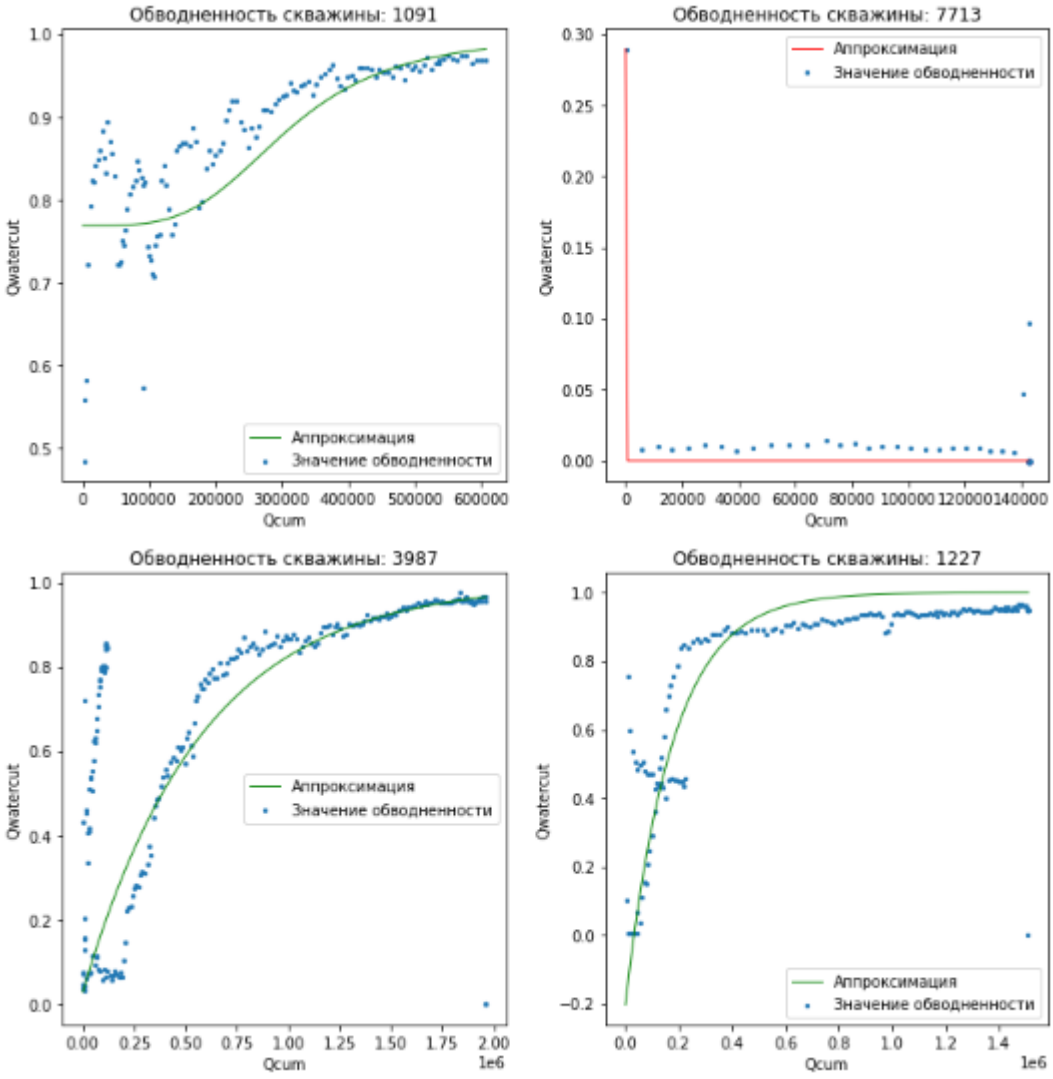

Рисунок 11. Графики обводненности скважин и соответствующие им аппроксимации.

На рисунках 12 и 13 представлено наложение всех аппроксимаций обводненностей скважин друг на друга для горизонтальных и наклонно-направленных скважин, соответственно. Можем наблюдать, что кривая обводненности, действительно, в большинстве случаев соответствует теории Баклея-Леверетта. Однако, также можно заметить, что в данном случае множество аппроксимаций наклонно-направленных скважин имеет большее сгущение в линейной части графика по сравнению с горизонтальными, что связано с количеством самих скважин в каждом наборе. В связи с этим, можем сделать вывод, что в случае с наклонно-направленными скважинами вероятно в результате кластеризации будет наблюдаться неравномерное заполнение кластеров с наибольшей наполненностью кластеров, находящихся вблизи линейной части графика.

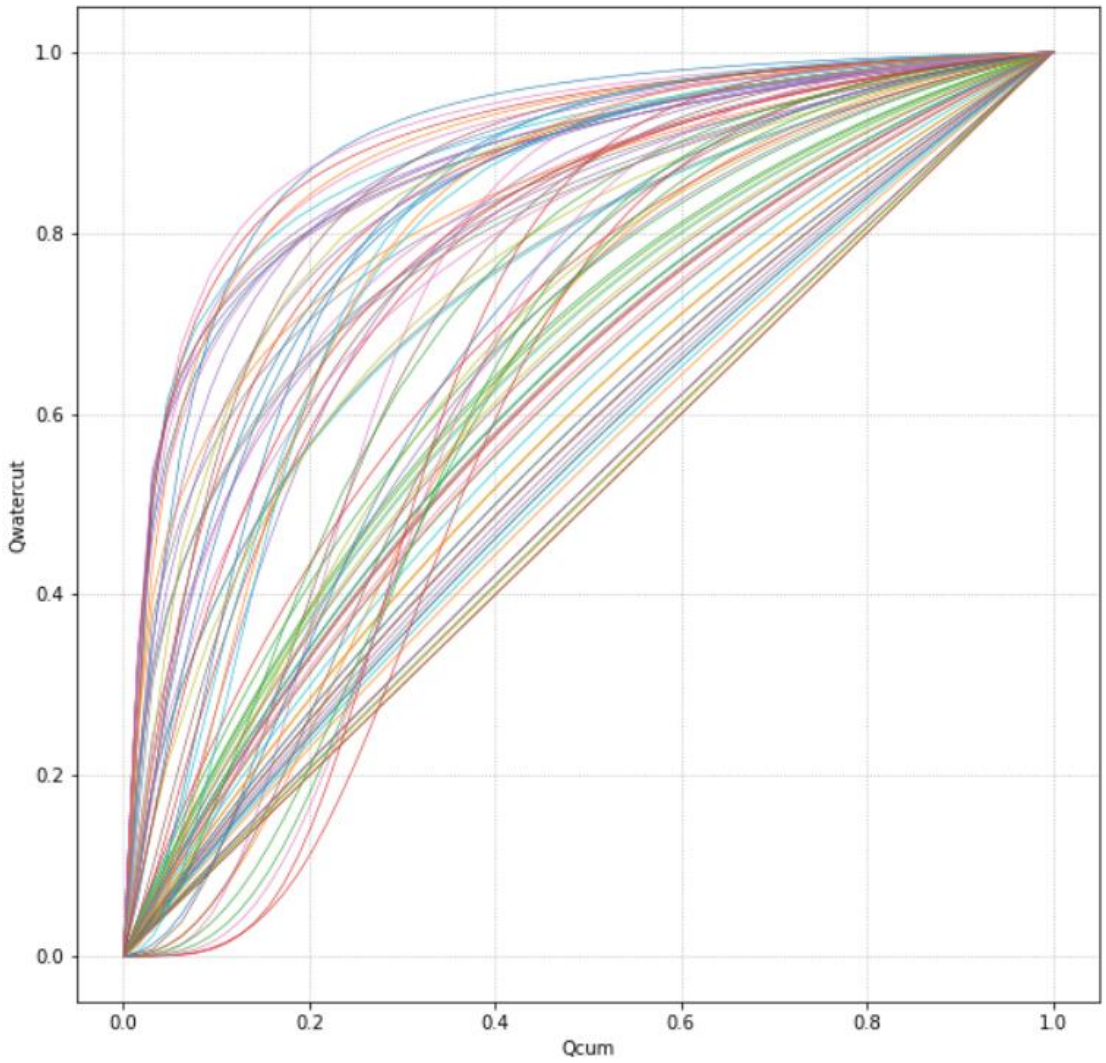

Рисунок 12. Наложение аппроксимаций по множеству горизонтальных скважин.

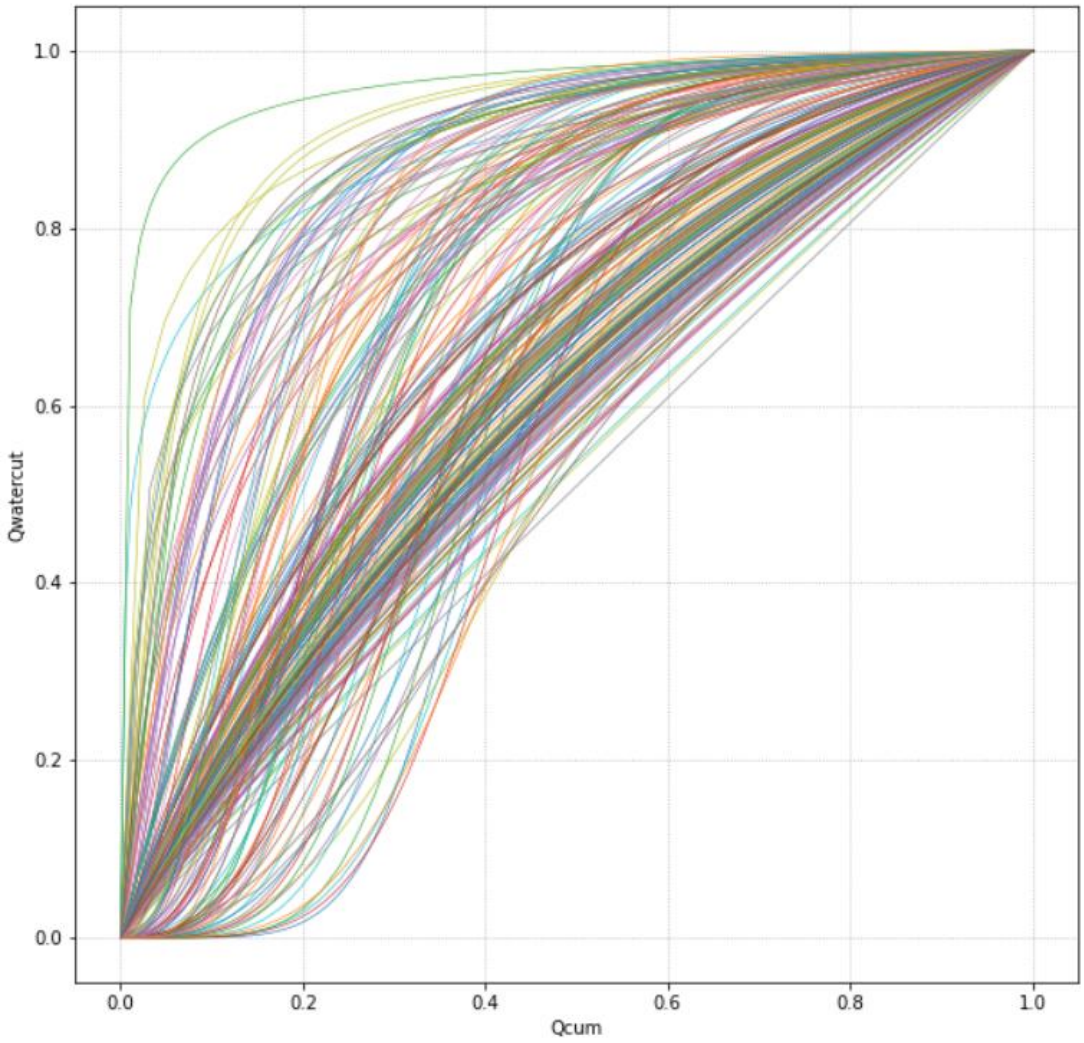

Рисунок 13. Наложение аппроксимаций по множеству наклонно-направленных скважин.

2.6. Задача кластеризации.

Кластеризация – одна из задач машинного обучения без учителя. В ходе кластеризации происходит разделение совокупности данных или отдельных точек данных на несколько групп таким образом, чтобы точки внутри одной группы были как можно более похожими на соседние и наиболее отличались от точек в других группах (кластерах). Например, точки на первом графике на рисунке 14 могут быть разделены на 3 кластера на основе оценки расстояний между любыми двумя соседними точками.

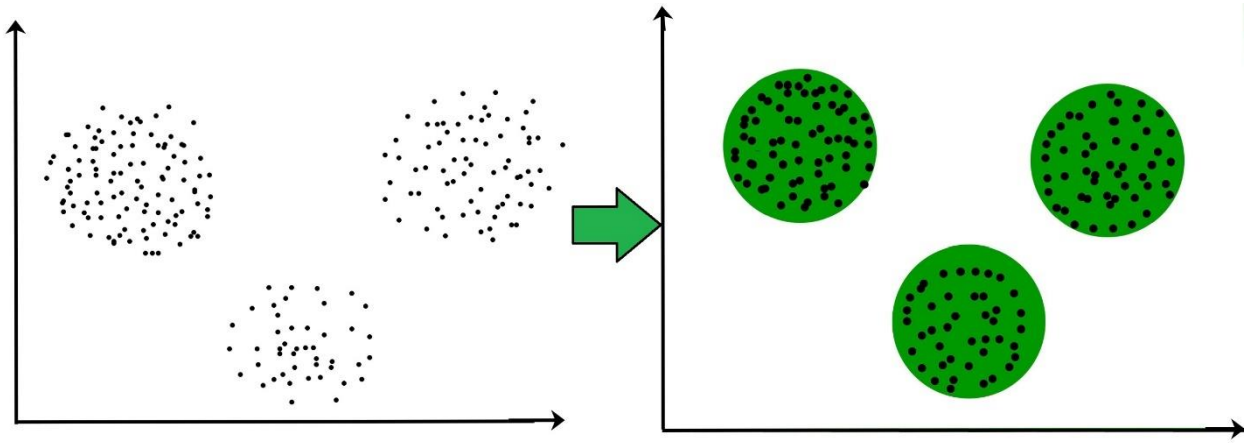

Рисунок 14. Пример кластеризации точек.

Под функциональной кластеризацией подразумевают ту же кластеризацию, но в которой множество точек описано функцией. Соответственно, при такой кластеризации кластер присваивается уже множеству функций. На рисунке 15 представлены примеры подобной кластеризации: пунктиром обозначены функции, а наиболее темной кривой сам кластер. Именно функциональная кластеризация и будет применяться далее.

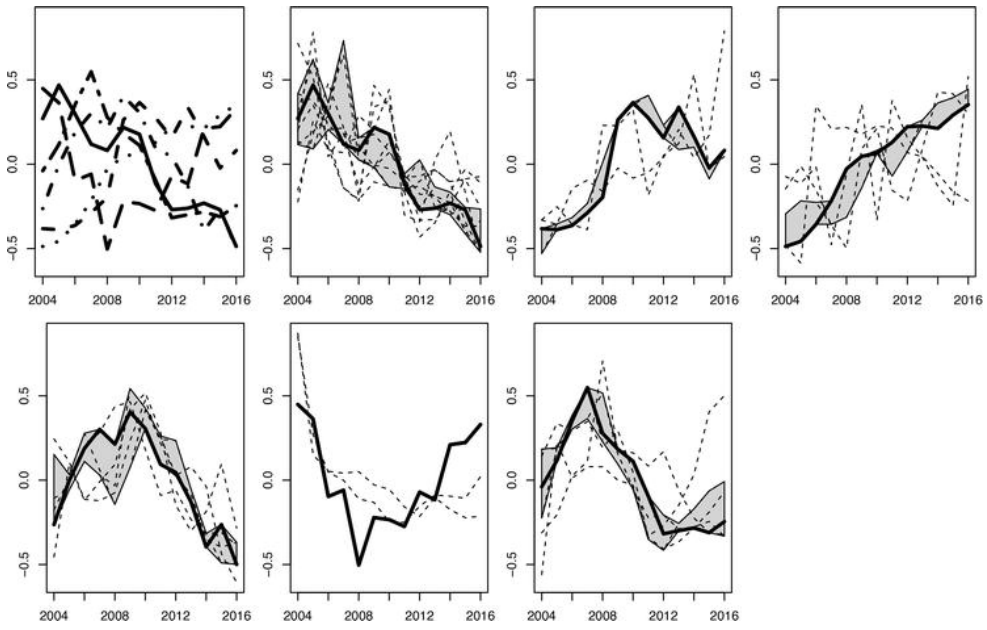

Рисунок 15. Пример функциональной кластеризации.

2.6.1. Метод k-средних

Метод k-средних – один из наиболее популярных методов кластерного анализа, целью которого является разделение m наблюдений на k кластеров, при этом каждое наблюдение присваивается тому кластеру, к центру (центроиду) которого оно ближе всего.

Пусть $X = \{x_1, x_2, \dots, x_n\}$ – вектор входных данных, где n – количество объектов. $V = \{v_1, v_2, \dots, v_c\}$ – вектор центров кластеров, определенных на X в пространстве размерности m, где c – количество кластеров, m – количество фичей, характеризующих исходные данные.

Цель метода – минимизировать функцию квадрата ошибки вида:

$$J_{KM}(X, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} D_{ij}^2$$

где D_{ij}^2 – квадрат выбранной меры расстояния

D_{ij} может быть любой p-нормой: $D_{ij} = \left|\left|x_{ij} - v_i\right|\right| = \left(\int_I \left|x_{ij} - v_i\right|^p dx\right)^{\frac{1}{p}}$, где I – область, в которой определен X, $1 \le i \le c, 1 \le j \le n_i, n_i$ – количество объектов в i-том кластере.

Для кластеров количества c метод k-средних с каждой итерацией минимизирует сумму расстояний от каждого наблюдения до его кластерного центроида. Каждая точка перемещается между кластерами до тех пор, пока сумма не может больше быть уменьшена. Алгоритм k-средних включает в себя следующие этапы:

1. Центроиды c кластеров выбираются из X;
2. Вычисляются расстояния между наблюдениями и центроидами;
3. Каждая точка назначается кластеру, центр которого наиболее близок к этой точек;
4. Центроиды кластеров обновляются по следующей формуле:

$$v_i = \sum_{i=1}^{n_i} \frac{x_{ij}}{n_i}, \qquad 1 \le i \le c$$

5. Расстояния от обновленных центроидов кластера пересчитываются;
6. Если ни одна точка не назначена новому кластеру, выполнение алгоритма останавливается, в противном случае повторяются шаги с 3 по 5 для вероятных перемещений точек между кластерами.

2.6.2. Агломеративный иерархический метод (АИМ)

Алгоритм агломеративной иерархической кластеризации является популярным примером методов кластеризаций, основанных на дендрограммах (рисунок)

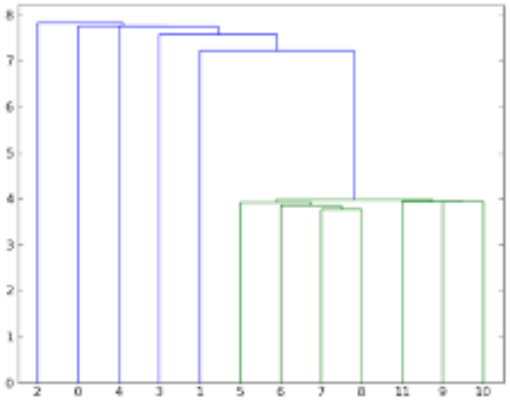

Рисунок 16. Пример дендрограммы.

Чтобы сгруппировать наборы данных в кластеры, используется подход "снизу вверх". Это означает, что этот алгоритм вначале рассматривает каждый набор данных как отдельный кластер, а затем начинает объединять ближайшую пару кластеров вместе. Он делает это до тех пор, пока все кластеры не будут объединены в один, содержащий все наборы данных. Эта иерархия кластеров и составляет дендрограмму. Рассмотрим принцип работы алгоритма:

1. Каждой точке присваивается кластер. Если существует N точек, то каждой из них будет присвоен свой кластер, общее количество кластеров тоже N (рисунок 17);

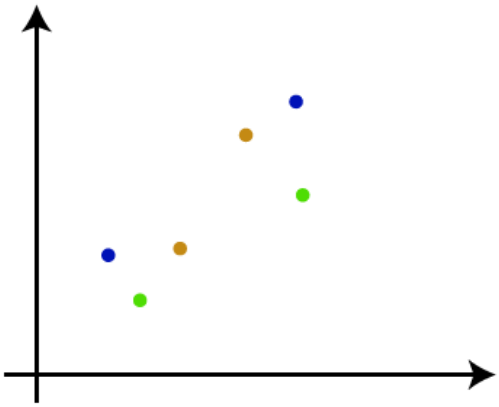

Рисунок 17. Первый шаг АИМ.

2. Берутся две ближайшие точки и объединяются в один кластер. Теперь количество кластеров становится $N - 1$ (рисунок 18);

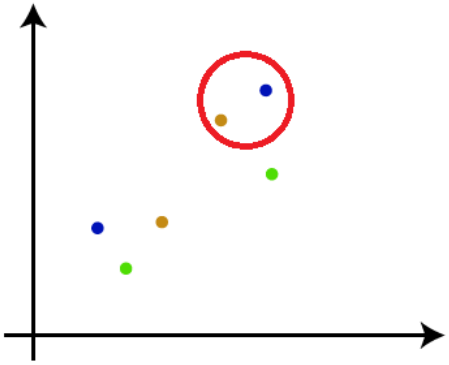

Рисунок 18. Второй шаг АИМ.

3. Снова берутся две ближайшие точки и образуют еще один кластер, количество кластеров становится $N - 2$ (рисунок 19);

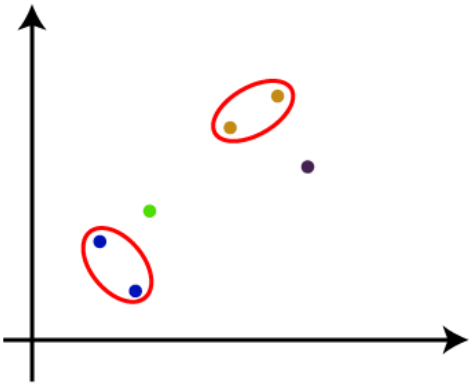

Рисунок 19. Третий шаг АИМ.

4. Третий шаг повторяется до тех пор, пока не останется один кластер (рисунок 20).

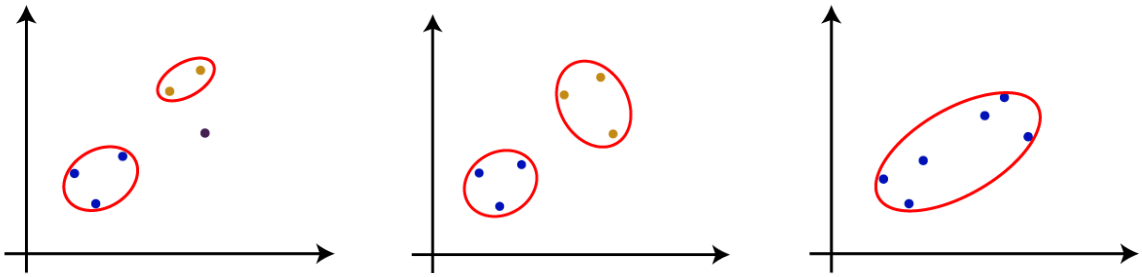

Рисунок 20. Последующие шаги АИМ.

5. Как только все кластеры сформировались в один большой кластер создается дендрограмма и необходимо по ней в обратную сторону спуститься до шага, на котором образовалось нужное количество кластеров.

2.7. Оценка качества кластеризации

В теоретическом плане проблема проверки адекватности кластеризации не решена, по крайней мере, без использования другого вида анализа или априорного знания принадлежности объектов к соответствующим классам [12]. В качестве метрики качества введем условие равномерной заполненности кластеров по проценту скважин в каждом из них, что теоретически должно повлиять на точность прогнозирования.

Сравним кластеризации скважин на основе 2–7 кластеров по методам из пунктов 2.6.1–2.6.2. Выбирать количество кластеров будем по наиболее оптимальной комбинации показателей отклонения медианной заполненности кластера от эталонной и волатильности по группам кластеров 2–4 и 4–7 из обоих методов кластеризации. Отклонение медианной заполненности кластера от эталонной будем считать более приоритетным, т.к. модели классификации показывают более точные результаты, когда тренировочные данные наиболее сбалансированы, т.е. имеют наиболее близкое количество элементов в каждом классе [1].

Под эталонной заполненностью кластера подразумевается процент, получаемый делением 100% на количество кластеров, а под волатильностью разница между максимальной и минимальной заполненностью кластера.

Таким образом, получим сравнительные таблицы 1–4. Для наибольшей интерпретируемости все значения представлены в процентах.

Количество кластеров	2	3	4	5	6	7
Эталонная заполненность кластера (%)	50	33.33	25	20	16.66	14.28
Медианная заполненность кластера (%)	50	33.33	25	17.59	17.59	17.59
Максимальная заполненность кластера (%)	64.81	35.19	32.41	29.63	21.30	18.52
Минимальная заполненность кластера (%)	35.19	31.48	17.59	12.96	10.19	6.48
Отклонение медианной заполненности от эталонной (%)	0	0	0	2.41	0.93	3.31
Волатильность (%)	29.62	3.71	14.82	16.67	11.11	12.04

Таблица 1. Горизонтальные скважины. Метод k-средних.

Количество кластеров	2	3	4	5	6	7
Эталонная заполненность кластера (%)	50	33.33	25	20	16.66	14.28
Медианная заполненность кластера (%)	50	35.19	19.44	15.74	12.97	15.74

Максимальная заполненность кластера (%)	64.81	45.37	45.37	45.37	45.37	24.07
Минимальная заполненность кластера (%)	35.19	19.44	15.74	6.48	2.78	2.78
Отклонение медианной заполненности от эталонной (%)	0	1.86	5.56	4.26	3.69	1.46
Волатильность (%)	29.62	25.93	29.63	38.89	42.59	21.29

Таблица 2. Горизонтальные скважины. АИМ.

Количество кластеров	2	3	4	5	6	7
Эталонная заполненность кластера (%)	50	33.33	25	20	16.66	14.28
Медианная заполненность кластера (%)	50	32.20	25.64	16.95	16.53	15.25
Максимальная заполненность кластера (%)	72.88	51.27	34.75	29.66	29.24	21.61
Минимальная заполненность кластера (%)	27.12	16.53	13.98	9.75	9.75	7.63
Отклонение медианной заполненности от эталонной (%)	0	1.13	0.64	3.05	0.13	0.97

Волатильность (%)	45.76	34.74	20.77	19.91	19.49	13.98

Таблица 3. Наклонно-направленные скважины. Метод k-средних.

Количество кластеров	2	3	4	5	6	7
Эталонная заполненность кластера (%)	50	33.33	25	20	16.66	14.28
Медианная заполненность кластера (%)	50	26.27	13.35	11.44	13.14	11.44
Максимальная заполненность кластера (%)	88.14	61.86	61.86	61.86	39.41	39.41
Минимальная заполненность кластера (%)	11.86	11.86	11.44	3.81	3.81	3.39
Отклонение медианной заполненности от эталонной (%)	0	7.06	11.65	8.56	3.52	2.84
Волатильность (%)	76.28	50	50.42	58.05	35.6	36.02

Таблица 4. Наклонно-направленные скважины. АИМ.

В результате кластеризации метод к-средних показал более оптимальные результаты по заданным метрикам в обоих случаях. В таблице 5 представлена итоговая выборка кластеров и методов кластеризации, которые далее будут использоваться в задаче прогнозирования.

Тип скважин	Горизонтальные		Наклонно-направленные	
Группы кластеров	2–4	5–7	2–4	5–7
Количество кластеров и метод	3 к-средних	6 к-средних	4 к-средних	6 к-средних

Таблица 5. Итоговая выборка кластеров и методов кластеризации.

На рисунках 21–24 визуализированы итоговые используемые выборки кластеров.

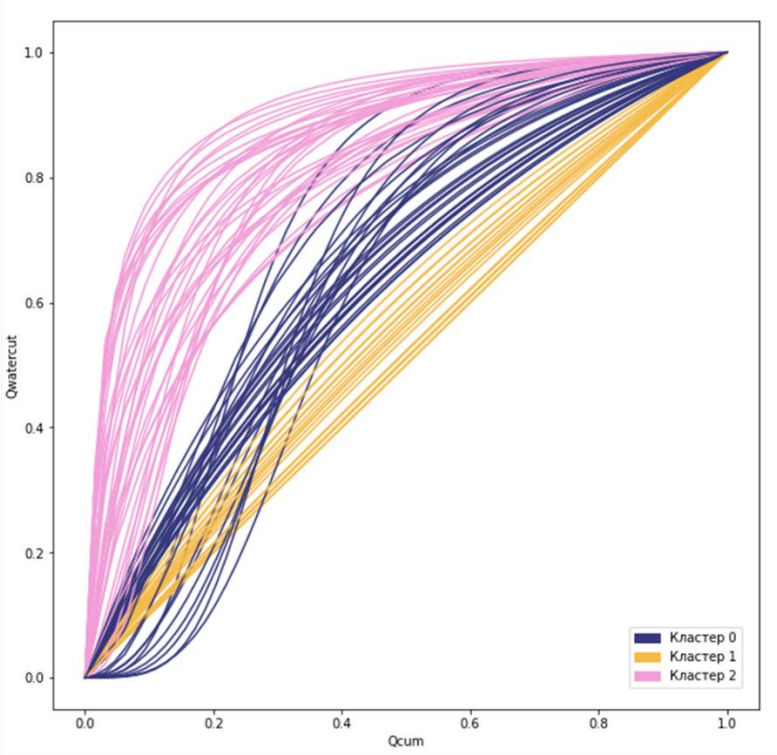

Рисунок 21. Горизонтальные скважины, 3 кластера, метод к-средних.

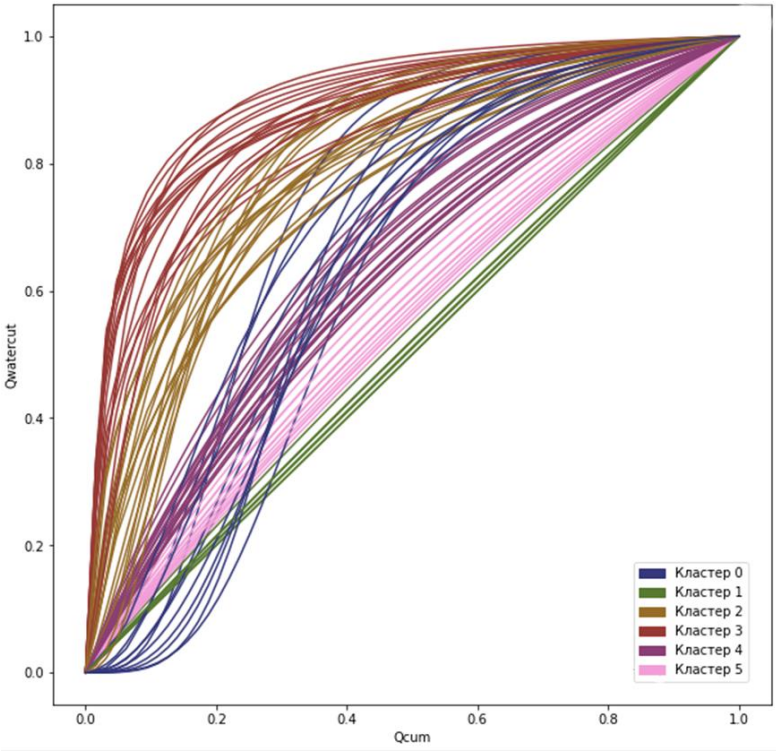

Рисунок 22. Горизонтальные скважины, 6 кластеров, метод к-средних.

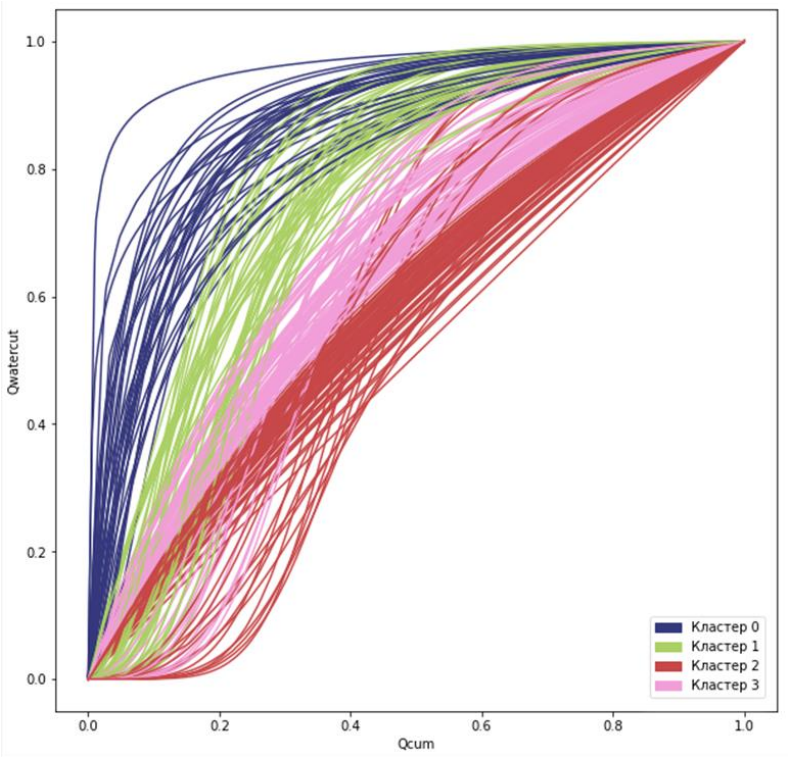

Рисунок 23. Наклонно-направленные скважины, 4 кластера, метод к-средних.

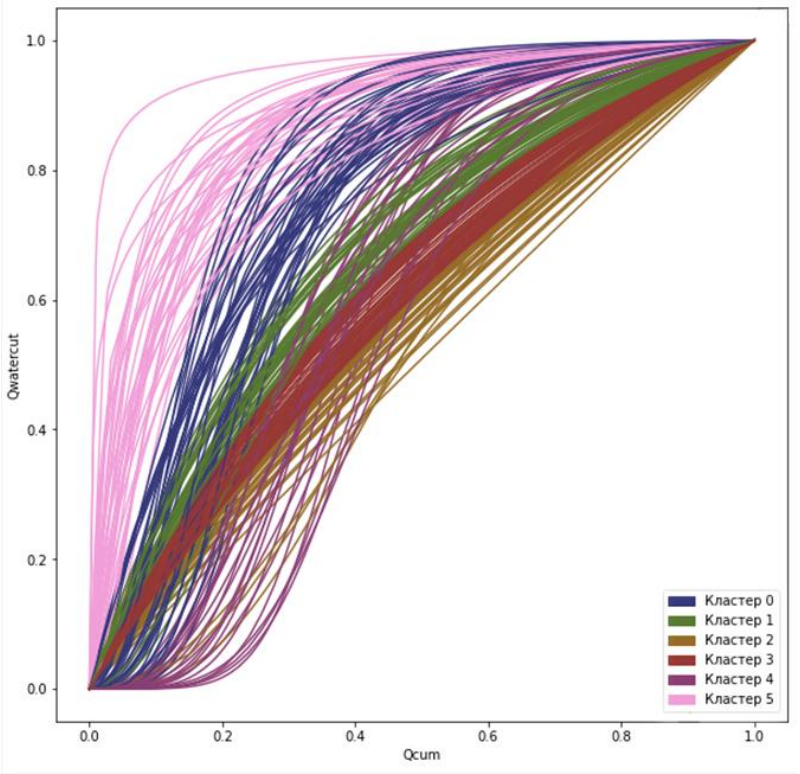

Рисунок 24. Наклонно-направленные скважины, 6 кластеров, метод к-средних.

ГЛАВА 3. ПРОГНОЗИРОВАНИЕ И ВАЛИДАЦИЯ МОДЕЛЕЙ.

3.1. Входные данные по геолого-физическим характеристикам

На рисунке 25 представлен вид исходной таблицы, дополненной данными скважин по их геолого-физическим характеристика. Представленные в ней колонки несут следующий смысл:

- скважина – идентификационный номер скважины;
- cluster – присвоенный скважине кластер;
- heff – эффективная мощность пласта;
- perm – проницаемость горных пород;
- kh – проводимость пласта;
- poro – пористость пласта.

Скважина	cluster	heff	perm	kh	poro
8436	2	12.907414	14.625736	188.780426	0.175354
4748	2	3.089955	136.194290	420.834281	0.202287
1338	2	7.299825	23.325754	170.273911	0.180139
8608	2	13.403497	25.298142	339.083562	0.180414
8192	2	12.098165	36.131580	437.125824	0.182114
...	...	...	...	...	...
5476	2	20.040877	17.095977	342.618376	0.162608
9156	1	9.437209	35.817438	338.016633	0.174700
7669	2	9.338336	32.640379	304.806833	0.179888
4710	2	26.124913	43.113849	1126.345561	0.179219
3358	2	7.236648	15.641657	113.193168	0.168544

Рисунок 25. Вид входных данных по геолого-физическим характеристикам скважин.

3.2. Обработка признаков

Масштабирование признаков – это одна из самых распространенных и сложных проблем в машинном обучении, но в то же время это одна из самых важных вещей, которую необходимо сделать правильно. Чтобы обучить модель машинного обучения и при этом получить наибольшую точность

нужно подать ей входные данные с известным набором признаков, диапазон значений которых должен быть либо увеличен, либо уменьшен до определенного отрезка в зависимости от обстоятельств. Таким образом, ключевая цель заключается в привидении различных данных в самых разных единицах измерения к единому виду, который позволит сравнивать их между собой или использовать для расчета схожести объектов.

Некоторые алгоритмы машинного обучения не требуют предварительного масштабирования, т.к. данная операция заключена в логике самого алгоритма, но для большинства алгоритмов данная операция является хорошей практикой, потому что позволяет алгоритму понять, что все признаки, условно говоря, равны между собой и никакому признаку не будет присвоен больший вес, чем другому [4].

Существует два распространенных способа масштабирования:

1. Нормализация (MinMax нормализация) – все значения масштабируются в указанном диапазоне, например от 0 до 1. Эта операция не влияет на распределение признаков, однако она усиливает последствия выбросов из-за меньших стандартных отклонений [4]. Поэтому выбросы необходимо устранить до нормализации. Имея на входе диапазон значений $[d_{min},\ d_{max}]$, нормализованное значение $x_{i\text{н}}$, соответствующее элементу x_i, лежащему на векторе X получается по следующей формуле:

$$x_{i\text{н}} = X_{\text{ст}}(d_{max} - d_{min}) + d_{min}$$

где $X_{\text{ст}} = \frac{x_i - X_{min}}{X_{max} - X_{min}}$ – стандартное отклонение,

X_{min}–минимальное значение элемента из X,

X_{max}–максимальное значение элемента из X.

2. Стандартизация (Z – оценка) – это операция масштабирования значений с учетом стандартного отклонения. Если стандартное отклонение признаков отличается, то и диапазон этих признаков будет отличаться. В результате влияние отклонений в признаках уменьшится [5]. Любая

точка x_i из X может быть приведена к стандартизированной форме x_{iz} по формуле:

$$x_{iz} = \frac{x_i - \mu}{\sigma}$$

где μ – математическое ожидание, σ – стандартное отклонение.

Стоит отметить, что для последующей интерпретации результатов машинного обучения используется обратный процесс масштабирования для приведения значений к исходному распределению.

Значения признаков строго положительные, поэтому применим нормализацию на диапазоне [0, 1]. В результате приведем признаки входных данных к виду, представленному на рисунке 26.

Скважина	cluster	heff	perm	kh	poro
4616	2	0.282535	0.077445	0.044076	0.380988
7907	2	0.000000	0.901269	0.105437	0.944016
5872	2	0.121155	0.136402	0.039182	0.481009
4916	2	0.296812	0.149768	0.083820	0.486771
6399	2	0.259246	0.223182	0.109745	0.522303
...	...	...	...	...	...
1135	2	0.487828	0.094185	0.084755	0.114525
1930	1	0.182667	0.221053	0.083538	0.367310
4165	2	0.179821	0.199524	0.074756	0.475765
5235	2	0.662920	0.270498	0.291994	0.461776
7564	2	0.119337	0.084330	0.024089	0.238622

Рисунок 26. Вид входных данных по геолого-физическим характеристикам скважин после нормализации.

3.3. Задача прогнозирования

Уточним, что под задачей прогнозирования, в данном случае, понимается задача мультиклассовой классификации с учителем, а не классическая задача прогнозирования будущих значений временного ряда на основе исторических данных.

В классической задаче классификации примеры классифицируются либо как 0, либо как 1. Такая классификация называется бинарной и задача алгоритма заключается в определении принадлежности примера к исследуемой группе или, наоборот, ее отсутствии. В мультиклассовой классификации, в отличии от бинарной, примеры классифицируются как принадлежащие к одному из ряда известных классов, количество которых больше 2 [18]. На рисунке 27 слева представлено разбиение примеров в соответствии с бинарной классификацией на 2 класса, а справа с мультиклассовой на 3 класса.

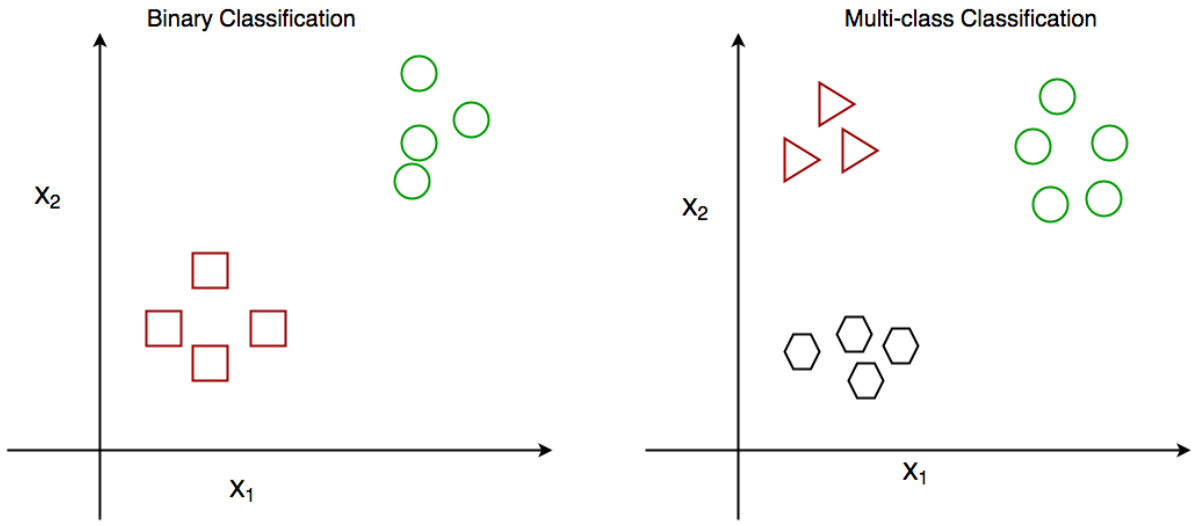

Рисунок 27. Сравнение бинарной (слева) и мультиклассовой классификации (справа).

В нашей задаче будем рассматривать именно мультиклассовую классификацию, классами которой будут служить присвоенные скважинам кластеры, а входными данными их геолого-физические характеристики. В результате по получаемым характеристикам скважины алгоритму необходимо определить принадлежность скважины к кластеру, который в свою очередь указывает на ожидаемую обводненность.

Отметим, что ввиду малого объема исходных данных рассматривать будем исключительно классические методы машинного обучения и их модификации.

3.3.1. Логистическая регрессия (One-vs-Rest)

Алгоритмы, разработанные для двоичной классификации, могут быть адаптированы для использования в мультиклассовых задачах. Для этого можно использовать стратегию подбора нескольких моделей бинарной классификации для каждого класса по сравнению с остальными классами (One-vs-Rest подход) или по одной модели для каждой пары классов (One-vs-One подход):

- One-vs-Rest: устанавливаем одну модель бинарной классификации для каждого класса против всех остальных классов;
- One-vs-One: устанавливаем одну модель бинарной классификации для каждой пары классов.

Итак, реализуем One-vs-Rest подход в комбинации с логистической регрессией. Опишем логику работы логистической регрессии на основе линейной регрессии.

В общем случае линейная модель делает прогноз, вычисляя взвешенную сумму признаков плюс константу, называемую смещением [7]:

$$y' = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

где y' – прогнозируемая величина, n – количество признаков,

x_i – значение i-го признака, θ_j – значение j-го параметра модели

(включая смещение θ_0 и веса признаков $\theta_1, \theta_2, \dots, \theta_n$)

В векторной форме можно представить так: $y' = h_\theta(x) = \theta \cdot x$, где θ – вектор весов модели, x – вектор значений признаков, h_θ – функция гипотезы.

Далее решается задача минимизации функционала квадратичной ошибки (MSE) на заданном векторе X:

$$MSE(X, h_\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

Таким образом, для перехода к логистической регрессии необходимо лишь взять логарифм от результата линейной регрессии. В векторной форме оценка вероятности логистической модели записывается так:

$$p' = h_\theta(x) = \sigma(x^T\theta), \qquad \sigma = \frac{1}{1+e^{-t}}$$

где σ – сигмоидная функция, которая выдает число от 0 до 1 (рисунок 28).

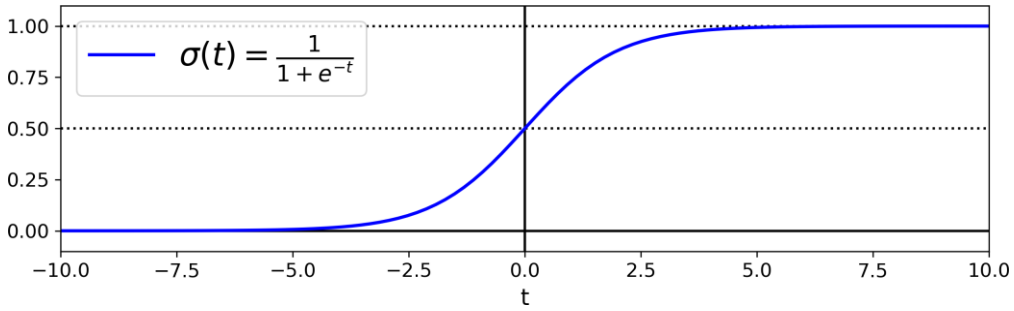

Рисунок 28. График сигмоиды.

Как только функция логистической регрессии оценивает вероятность p' того, что значение x принадлежит к какому-либо классу, так сразу возвращается прогнозируемое значение y':

$$y' = \begin{cases} 0 \text{ при } p' < 0.5 \\ 1 \text{ при } p' \geq 0.5 \end{cases}$$

Заметим, что $\sigma(t) < 0.5$, когда $t < 0$, и $\sigma(t) \geq 0.5$, когда $t \geq 0$, т.е. логистическая регрессия выдает 1, когда $x^T\theta$ положительное, и 0, когда отрицательное.

3.3.2. Random forest

Алгоритмы, основанные на решающих деревьях хоть и имеют простую логику, но тем не менее являются мощными решениями и демонстрируют высокую точность в определенных случаях. Метод случайного леса (Random forest) основан на сочетании идей бэггинга Брейнмана [7] и метода случайных подпространств [8]. Идея метода заключается в построении ансамбля решающих деревьев, каждое из которых не имеет большую предиктивную способность, но при комбинировании большого количества результирующий алгоритм обладает высокой точностью. Пример дерева решений представлен на рисунке 29.

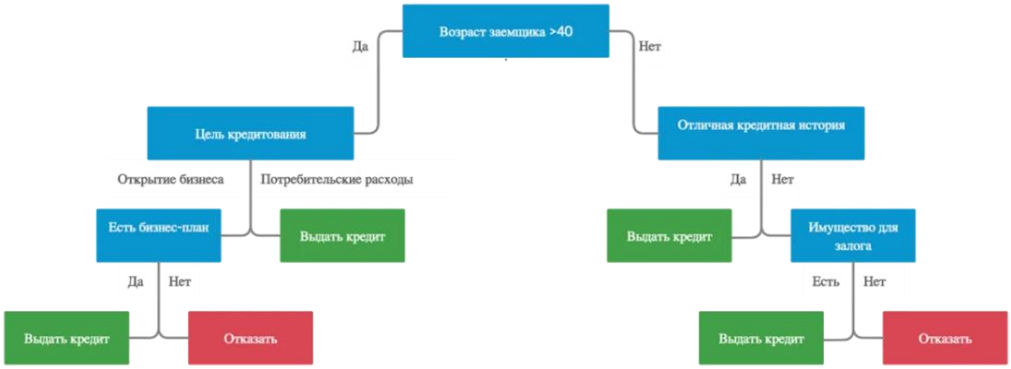

Рисунок 29. Пример дерева решений.

Алгоритм случайного леса, состоящий из N деревьев на основе обучающей выборки X выглядит следующим образом:

1. Для каждого из N деревьев создается выборка X_n, где $n \in (1, \dots, N)$ с помощью бутстрэпа – из всего множества обучающей выборки равновероятно выбирается заданное число объектов с возвращением, то есть один и тот же объект может быть использован в выборке несколько раз;
2. Строится ответ решающего дерева для текущей подвыборки $t_n(X_n)$;
3. После получения ответа для каждого дерева ответ итогового алгоритма находится как среднее арифметическое:

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} t_i(x)$$

3.3.3. Extra-Trees

Алгоритм экстремально случайных деревьев (Extra-Trees) является модификацией алгоритма случайных деревьев. Спускаясь по дереву в алгоритме случайных деревьев, в каждом узле для разделения рассматривается только случайное подмножество объектов. Деревья можно сделать еще более случайными, если использовать случайные пороговые значения для каждого объекта, а не ища наилучшие возможные, как это происходит в обычных деревьях решений. Это приводит к более низкой дисперсии и позволяет обучать дополнительные деревья намного быстрее, чем в алгоритме случайного леса, поскольку поиск наилучшего возможного порога в каждом

узле – одна из самых трудоемких задач в алгоритме случайного леса [23].

3.3.4. Градиентный бустинг

Как и алгоритм случайный леса, градиентный бустинг представляет собой композицию (ансамбль) простых алгоритмов, чаще всего деревьев решений. Градиентный бустинг работает путем последовательного добавления новых моделей в ансамбль, каждая из которых постепенно улучшает результат своего предшественника. Алгоритм работы градиентного бустинга следующий:

1. Построить простой алгоритм: $b_0(x)$ и внести его в ансамбль $a_0(x) = b_0(x)$
2. На каждом шаге итерации вычислить вектор сдвига:

$$s = -\nabla F = \begin{pmatrix} -L'_z\big(y_1, a_{n-1}(x_1)\big) \\ \dots \\ -L'_z\big(y_l, a_{n-1}(x_l)\big) \end{pmatrix}$$

   И построить новую модель:

$$b_n(x) = argmin_b \frac{1}{l}\sum_{i=1}^{l}(b(x_i) - s_i)^2$$

   Необходимо подобрать параметры так, чтобы значения алгоритма на обучающей выборке были как можно ближе к вычисленному вектору сдвига.
3. Остановить итерационный процесс при выполнении критерия остановки.

3.4. Метрики качества

Основными метриками в задачах классификации являются accuracy, precision и recall. Accuracy отвечает за долю правильных ответов алгоритма. Accuracy может не отображать реальную ситуацию в случае задач с несбалансированными классами, как наша. В данном случае она не различает разные виды ошибок. Precision и recall являются метриками, которые оценивают качество работы алгоритмов по отдельности на каждом из классов. Точность (precision) отвечает за то, какая часть объектов, которым алгоритм

классификации присвоил класс, действительно принадлежит ему. Полнота (recall) дает понять, какую часть объектов указанного класса из всех нашел алгоритм.

На рисунке 30 представлена графическая интерпретация метрик precision и recall.

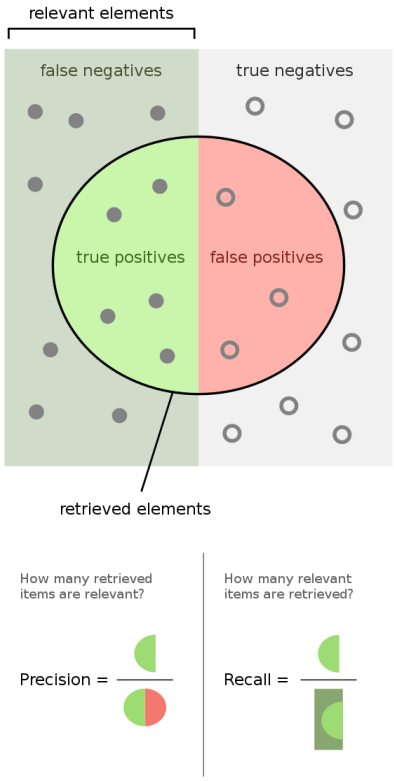

Рисунок 30. Точность и полнота.

Иногда недостаточно использования только одной метрики и требуется получить их обобщение. В таком случае используется агрегированный критерий, который называется F-мера:

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

где β – отвечает за весь precision в данном критерии. В данной работе будет использоваться F1-мера, где β=1. Значения данного критерия принадлежат промежутку от 0 до 1. Чем больше precision и recall, тем ближе F-мера к единице.

3.5. Сравнение результатов машинного обучения

Тренировочная выборка составит 80% от всех данных по обоим скважинам, на остальных 20% будем замерять точность по средней межкластерной F-мере. В таблицах 6–7 представлено сравнение результатов машинного обучения для горизонтальных и наклонно-направленных скважин, соответственно, в соответствии со следующей нумерацией алгоритмов:

1. Логистическая регрессия;
2. Random forest;
3. Extra-Trees;
4. Градиентный бустинг.

Количество кластеров	3				6			
Алгоритм	1	2	3	4	1	2	3	4
F-мера	0.45	0.61	0.66	0.46	0.19	0.38	0.45	0.21

Таблица 6. Результаты машинного обучения для горизонтальных скважин.

Количество кластеров	4				6			
Алгоритм	1	2	3	4	1	2	3	4
F-мера	0.34	0.51	0.42	0.37	0.27	0.31	0.30	0.20

Таблица 7. Результаты машинного обучения для наклонно-направленных скважин.

Для горизонтальных скважин с разделением на 3 и 6 кластера лучшую точность показал алгоритм Extra-Trees с точностью в 0.66 и 0.45, соответственно. Для наклонно-направленных скважин с разделением на 4 и 6 кластера лучшую точность показал алгоритм Random forest с точностью в 0.51 и 0.31, соответственно.

Как было сказано ранее, можем наблюдать, что из-за меньшей степени сбалансированности кластеров результаты для наклонно-направленных скважин в целом получились хуже, несмотря на то, что количество данных по ним больше, чем по горизонтальным скважинам. Заметим также, что с увеличением количества кластеров точность уменьшается, что связано как с ухудшением сбалансированности кластеров, так и с повышением сложности расчетов для моделей ввиду необходимости определения большего количества классов.

ЗАКЛЮЧЕНИЕ

В ходе данной работы была проведена очистка и фильтрация исходных данных по множеству скважин. Был проведен анализ шести алгоритмов аппроксимации данных и оптимизирован с помощью нелинейного метода наименьших квадратов, а точность каждой аппроксимации была оценена с помощью средней абсолютной ошибки, после чего выбиралась наиболее точная аппроксимация и сразу же принималось решение о том оставлять скважину с этой аппроксимацией или нет в зависимости от наклона кривой.

Затем была проведена кластеризация всего множества аппроксимаций двумя методами по 2–7 кластерам, была проведена оценка заполненности кластеров, на основании которой принималось решение о выборе наиболее оптимального количества кластеров из групп, содержащих 2–4 и 5–7 кластеров. Как результат, оптимальными оказались разбиения на 3 и 6 кластеров для горизонтальных скважин, и на 4 и 6 кластеров для наклонно-направленных скважин.

Наконец, на основе размеченных по кластерам скважин было произведено прогнозирование с помощью четырех классических и модифицированных методов машинного обучения путем решения задачи мультиклассовой классификации, в ходе которого было заключено, что наиболее точными показал себя метод Extra-Trees при разбиении на 3 и 6 кластеров для горизонтальных скважин, метод Random forest при разбиении на 4 и 6 кластеров для наклонно-направленных скважин.

Таким образом, можно считать, что цель выпускной квалификационной работы выполнена и поставленные задачи решены полностью.

В дальнейшем возможны следующие направления развития работы для повышения точности:

- Увеличение объема входных данных по скважинам, что позволит отфильтровать множество скважин таким образом, что кластеры имеют схожую наполненность. Все это позволит применить более сложные и эффективные алгоритмы машинного обучения;

- Сравнение как большего количества методов аппроксимации, кластеризации и классификации, так и более глубокой и тонкой настройки их гиперпараметров;
- Проведение аналитического исследования, включающего в себя поточечную проработку и валидацию входных данных;
- Увеличение количества признаков как напрямую, т.е. путем сбора дополнительной информации о геолого-физических характеристиках скважин, так и косвенно, т.е. путем проверки различных предположений о корреляции обводненности со сторонними факторами, такими как простои оборудования, сменами работников, экономической и финансовой частью добычи нефти, условиями погоды.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Шитиков В. К., Мастицкий С. Э., Классификация, Регрессия и Другие Алгоритмы Data Mining с Использованием R, Оценка качества Кластеризации – 2017.
2. Apostolos Kantzas, Jonathan Bryan, Saeed Taheri, Fundamentals of Fluid Flow in Porous Media, Chapter 4, Immiscible Displacement, Buckley-Leverett Theory // Perm Inc. – 2018.
3. El-Gohary A., Alshamrani Ahmad, Al-Otaibi Adel Naif, The Generalized Gompertz Distribution // Volume 37, Issues 1-2, Elsevier. – 2013. – P. 13-24.
4. Géron Aurélien, Hands on Machine Learning with Scikit Learn, Keras and Tensorflow Concepts, Tools and Techniques to Build Intelligent Systems // O'Reilly Media – 2019 – P. 102-104
5. Géron Aurélien, Hands on Machine Learning with Scikit Learn, Keras and Tensorflow Concepts, Tools and Techniques to Build Intelligent Systems // O'Reilly Media – 2019 – P. 114-115
6. Géron Aurélien, Hands on Machine Learning with Scikit Learn, Keras and Tensorflow Concepts, Tools and Techniques to Build Intelligent Systems // O'Reilly Media – 2019 – P. 200
7. Breiman. L, Bagging predictors, Machine Learning. 24 (2): 123–140, 1996.
8. Ho, Tin Kam, The Random Subspace Method for Constructing Decision Forests, 1998.
9. Gove Jeffrey H., Lynch Thomas B., Ducey Mark J., The Chapman–Richards Distribution and its Realationship to the Generalized Beta // Forest Ecosystems. – 2019.
10. Madsen K., Nielsen H.B., Tingleff O. Methods for non-linear least squares problems // Technical University of Denmark (IMM). - 2004. – P. 1-2.
11. Damcese M.A.EL, Abdelfattah Mustafa, Eliwa M.S., Exponential Generalized Weibull – Gompertz Distribution // Tanta University, Mansoura University. – 2014.

12. Nazri Mohd Nawi, Walid Hasen Atomi, M.Z. Rehman, The Effect of Data Preprocessing on Optimized Training of Artificial Neural Networks // Procedia Technolagy, Volume 13. – 2013. – P. 32-39.
13. ansdimat.com, Основные Фильтрационные Параметры Пласта, Проводимость, URL: http://ansdimat.com/ru/param.shtml#. Дата обращения: 20.05.2022.
14. IBM, Cloud Education, Machine Learning, URL: https://www.ibm.com/cloud/learn/machine-learning. Дата обращения: 20.05.2022.
15. IBM, Prepare Your Data for AI and Data Science, URL: https://www.ibm.com/garage/method/practices/code/data-preparation-ai-data-science/. Дата обращения: 30.05.2022.
16. neftegaz.ru, Наклонно-направленное Бурение. URL: https://neftegaz.ru/tech-library/tekhnologii/141796-naklonno-napravlennoe-burenie/. Дата обращения: 20.05.2022.
17. neftegaz.ru, Наклонно-направленное Бурение. URL: https://neftegaz.ru/tech-library/tekhnologii/141796-naklonno-napravlennoe-burenie/. Дата обращения: 20.05.2022.
18. neftegaz.ru, Пористость. URL: https://neftegaz.ru/tech-library/ngk/147694-poristost/. Дата обращения: 20.05.2022.
19. neftegaz.ru, Проницаемость Горных Пород Пласта, URL: https://neftegaz.ru/tech-library/geologorazvedka-i-geologorazvedochnoe-oborudovanie/141979-pronitsaemost-gornykh-porod-plasta. Дата обращения: 20.05.2022.
20. neft.academic.ru, Нефтегазовая микроэнциклопедия, Мощность пласта, URL: https://neft.academic.ru/213/Мощность_пласта. Дата обращения: 20.05.2022.
21. oilloot.ru, Разработка Нефтяных и Газовых Месторождений, Задача Баклея-Леверетта и ее обобщения, URL: http://oilloot.ru/77-geologiya-

geofizika-razrabotka-neftyanykh-i-gazovykh-mestorozhdenij/271-zadacha-bakleya-leveretta-i-ee-obobshcheniya. Дата обращения: 22.05.2022.

22. towardsdatascience.com, Vinicius Trevisan, Comparing Robustness of MAE, MSE and RMSE, URL: https://towardsdatascience.com/comparing-robustness-of-mae-mse-and-rmse-6d69da870828. Дата обращения: 30.05.2022.