**Аналитическая записка:**

**«Модель машинного обучения для прогнозирования запасов газа по кривым ГИС»**



Работу выполнил

**Ильин Илья**

Содержание

[Введение 3](#_Toc184895622)

[Глава 1. Обзор стандартных методов расчета фильтрационно-емкостных свойств 4](#_Toc184895623)

[1.1 Основные методы оценки пористости по ГИС 4](#_Toc184895624)

[Комплексный метод расчета пористости: 4](#_Toc184895625)

[1.2 Ограничение стандартных методов 5](#_Toc184895626)

[1.3 Оценка качества расчета пористости по ГИС 6](#_Toc184895627)

[Глава 2. Реализация алгоритма машинного обучения для прогноза фильтрационно-емкостных свойств горных пород 9](#_Toc184895628)

[2.1 Перспективы использования машинного обучения в повышении качества прогнозирования по кривым ГИС 9](#_Toc184895629)

[2.2 Подготовка исходных данных для машинного обучения. 11](#_Toc184895630)

[2.3. Разведочный анализ данных (EDA). 11](#_Toc184895631)

[2.4. Выбор модели для обучения 23](#_Toc184895632)

[2.5. Подбор гиперпараметров для усовершенствования модели Random Forest. 27](#_Toc184895633)

[2.6. Оценка модели на различных размерах тестовой выборки 31](#_Toc184895634)

[2.7. Оценка применимости результатов моделирования в производстве. 32](#_Toc184895635)

[2.8. Перспективы использования машинного обучения в нефтегазовой геологии. 36](#_Toc184895636)

[Заключение 38](#_Toc184895637)

[Список литературы 40](#_Toc184895638)

# Введение

Основным параметром ФЕС горных пород залежей нефти и газа является пористость. Данный параметр напрямую влияет на величину запасов углеводородов, а значит обосновывает экономическую перспективу проекта.

Целью настоящей работы является повышение качества и скорости прогнозирования фильтрационно-емкостных свойств горных пород месторождений углеводородов на основе машинного обучения.

Для повышения точности оценки пористости целесообразно использовать комплексный подход с учетом нескольких каротажных методов и обязательной калибровкой по керну.

Применение машинного обучения для прогнозирования пористости и других параметров по ГИС может значительно повысить точность и надежность расчетов, особенно при интеграции с данными керна и учетах поправок на влияние глинистости и насыщенности.

Для достижения цели необходимо выполнить следующие задачи:

- подготовить набор данных для машинного обучения,

- выполнить разведочный анализ данных, определив их статистические характеристики и применимость для обучения модели;

- выполнить подготовку признаков для машинного обучения;

- подобрать метрики оценки результатов работ с учетом специфики целевой переменной;

- построить несколько моделей машинного обучения, выбрать наиболее подходящую модель;

- усилить выбранную модель машинного обучения и выполнить оценку прогноза для применимости использования модели в производственном процессе.

# Глава 1. Обзор стандартных методов расчета фильтрационно-емкостных свойств

Основным параметром ФЕС горных пород залежей нефти и газа является пористость. Данный параметр напрямую влияет на величину запасов углеводородов, а значит обосновывает экономическую перспективу проекта. Пористость по данным геофизических исследований скважин (ГИС) для коллекторов нефти и газа рассчитывается с использованием различных методов и типов каротажных кривых, таких как нейтронный, акустический, гамма-гамма-каротаж и другие [1, 2, 3]. На практике пористость обычно рассчитывается с применением эмпирических или полуэмпирических уравнений, основанных на калибровке кривых ГИС с данными керна и типом породы [2, 4] .

### **1.1 Основные методы оценки пористости по ГИС**

**По акустическому каротажу (АК):** Акустический каротаж измеряет скорость прохождения звуковой волны через горную породу, что позволяет оценить пористость. Формула для расчёта пористости по АК основана на законе Вилля [1].

**По гамма-гамма-каротажу (ГГК):** Гамма-гамма-каротаж измеряет плотность породы, из которой также можно рассчитать пористость. Формула пористости по ГГК [3].

**По нейтронному каротажу**: Нейтронный каротаж чувствителен к содержанию водорода в порах породы, поэтому его данные напрямую используют для оценки пористости в водонасыщенных коллекторах. При наличии газа результаты могут быть скорректированы [2, 3].

### Комплексный метод расчета пористости**:**

Для точности рекомендуется использовать комбинацию каротажей: акустического, гамма-гамма и нейтронного. Комплексный метод позволяет учесть возможные отклонения в измерениях и точнее рассчитать пористость. Например, в газонасыщенных коллекторах нейтронный и плотностной каротаж сочетают для устранения погрешности из-за низкого содержания водорода [2, 3].

Формулы и подходы могут меняться в зависимости от типа породы и насыщения [4].

**1.2 Ограничение стандартных методов**

Использование данных ГИС для расчета пористости — широко распространённый метод, однако он имеет несколько слабых сторон:

**Ограниченная точность при неоднородной минералогии:**

Формулы для расчета пористости (например, по плотностному или акустическому каротажу) предполагают определенную минералогическую однородность коллектора (например, песчаник или известняк). В условиях сложной, неоднородной минералогии (например, глинистых песчаников или доломитизированных известняков) данные ГИС могут давать неточные значения пористости [2, 4].

Неоднородная минералогия вызывает изменчивость плотности матрицы или скорости звука, что требует введения поправок, а их определение увеличивает трудозатраты и возможность ошибок.

**Влияние флюидного состава:**

ГИС плохо различают флюиды, особенно в газонасыщенных коллекторах, где нейтронный и плотностной каротаж могут давать завышенные или заниженные значения пористости.

Если в порах содержатся не только вода и нефть, но и газ, это может значительно повлиять на показания нейтронного каротажа и плотностного каротажа, приводя к неточным расчетам. Использование корректирующих коэффициентов требует детальной калибровки [3, 4].

**Чувствительность к глинистости:**

Глинистость коллектора может серьезно искажать результаты измерений, особенно в нейтронном и акустическом каротаже. Глина удерживает воду на молекулярном уровне, что влияет на оценку пористости, завышая ее.

Для учета влияния глинистости нужно либо применять эмпирические коррекции, либо использовать дополнительные исследования, что увеличивает сложность обработки данных и снижает точность.

**Необходимость калибровки по данным керна:**

Для получения точных значений пористости требуется калибровка данных ГИС по лабораторным данным керна. Это особенно важно в коллекторах со сложной пористо-проницаемой структурой, но не всегда возможно из-за отсутствия кернового материала.

Без калибровки данные ГИС могут давать лишь приблизительные значения, что снижает надежность оценки пористости.

**Чувствительность к качеству каротажа и наличию помех:**

Любые нарушения в процессе каротажного измерения, такие как плохой контакт зонда с породой, плохие условия заполнения скважины (например, наличие бурового раствора), могут приводить к погрешностям.

Шумы и помехи в сигнале могут существенно исказить измеренные значения, особенно на больших глубинах и в высокопористых породах, где погрешности могут усиливаться.

Для повышения точности оценки пористости целесообразно использовать комплексный подход с учетом нескольких каротажных методов и обязательной калибровкой по керну.

## 1.3 Оценка качества расчета пористости по ГИС

Оценка качества расчета пористости по данным геофизических исследований скважин (ГИС) проводится по нескольким основным критериям.

1. **Сравнение с лабораторными данными керна.**

**Калибровка и проверка:** наиболее точный метод — это сравнение расчетной пористости по ГИС с результатами лабораторных измерений пористости керна. Обычно данные пористости из ГИС корректируют с помощью данных керна для повышения точности, и если разница между ними мала, расчет считается качественным [3, 4].

**Анализ расхождений**: Строят графики зависимости пористости по ГИС от керновой пористости и анализируют отклонения. Допустимое расхождение для разных типов пород обычно составляет 2-3%.

2. **Согласованность между различными методами ГИС**

**Кросс-проверка методов**: Пористость часто оценивается по нескольким методам (плотностной, нейтронный и акустический каротажи), и качество расчетов можно оценить по их согласованности. Значительные расхождения между методами могут свидетельствовать о необходимости корректировок [1].

**Анализ графиков**: например, на графике нейтрон-плотностного каротажа оценка качества зависит от положения точек по отношению к линии соответствия пористости и насыщенности. Расхождения обычно указывают на ошибки в измерениях или на специфические особенности породы.

3. **Качество обработки и наличие корреляций с другими параметрами**

**Плотностные и акустические свойства**: если значения пористости по ГИС находятся в разумной зависимости с другими физическими параметрами (например, плотностью, скоростью звука), это указывает на надежность оценки [2].

**Наличие аномалий**: если расчеты показывают аномальные значения, это может быть признаком некорректной калибровки или помех при измерении, например, из-за глинистости или наличия газа.

4. **Проверка влияния факторов, осложняющих расчет**

**Глинистость**: Глинистые включения могут давать завышенные значения пористости, особенно в нейтронном каротаже, поэтому применяют корректировки, а затем проверяют качество расчетов на основе скорректированных значений.

**Насыщенность флюидами**: Пористость может быть завышена в газонасыщенных зонах. Поэтому после учета флюидного состава оценивают пористость, проверяя, соответствуют ли данные по насыщенности общим характеристикам коллектора.

5. **Статистическая оценка качества расчетов**

**Коэффициенты корреляции**: рассчитывают коэффициенты корреляции между пористостью по ГИС и керновыми данными. Высокие значения корреляции (обычно выше 0,8) указывают на хорошее качество расчета.

**Среднеквадратическая ошибка (RMSE)**: используют для количественного измерения расхождений между расчетными и фактическими значениями пористости. Чем меньше RMSE, тем выше качество [4].

6. **Визуальные методы анализа**

**Гистограммы и распределения**: Гистограммы пористости позволяют оценить наличие аномальных значений, выявить смещения и оценить корректность расчетов.

**Кроссплоты**: Построение кроссплотов (например, плотностного и нейтронного каротажей) позволяет визуально оценить соответствие расчетной пористости стандартным тенденциям и выявить отклонения, которые могут свидетельствовать о проблемах в расчете.

Для комплексной оценки качества расчета пористости необходимо учитывать все вышеуказанные аспекты, что позволяет выявить и скорректировать возможные погрешности и получить наиболее точные результаты.

# Глава 2. Реализация алгоритма машинного обучения для прогноза фильтрационно-емкостных свойств горных пород

## 2.1 Перспективы использования машинного обучения в повышении качества прогнозирования по кривым ГИС

Машинное обучение (МЛ) может значительно повысить качество прогнозирования пористости и других параметров по данным геофизических исследований скважин (ГИС) [5, 6].

1. **Учет нелинейных зависимостей**

* Данные ГИС включают множество параметров (например, плотностной, нейтронный и акустический каротажи), которые нелинейно взаимосвязаны с пористостью и другими свойствами коллектора [7]. МЛ-алгоритмы, такие как градиентный бустинг, случайный лес или нейронные сети, хорошо работают с нелинейностями и могут более точно предсказать параметры, чем линейные модели [8].
* Использование ансамблевых методов (например, случайного леса и бустинга) также помогает учесть сложные нелинейные взаимосвязи, повышая точность прогноза пористости и других свойств.

2. **Обработка большого объема данных**

* Современные месторождения генерируют огромные объемы ГИС-данных, особенно в рамках высокоразрешающих исследований. МЛ-алгоритмы могут эффективно работать с большими объемами данных, выявляя скрытые закономерности и повышая точность прогнозов [6].

3. **Калибровка с использованием лабораторных данных**

* МЛ помогает корректировать данные ГИС с использованием лабораторных данных по керну, обучая модель на этих данных и получая точные прогнозы пористости в скважинах без керновых образцов. Таким образом, можно обобщить зависимости, найденные на керне, на другие участки месторождения [9].

4. **Коррекция влияния глинистости и насыщенности флюидами**

* Машинное обучение позволяет учитывать влияние глинистости, влажности и других факторов, которые осложняют интерпретацию данных ГИС. Например, алгоритмы могут обучиться автоматически корректировать завышенные значения пористости в зонах с высоким содержанием глинистых частиц или в газонасыщенных зонах [8].
* При наличии данных о насыщенности скважин можно также использовать алгоритмы МЛ для прогнозирования этих параметров и их коррекции на расчет пористости.

5. **Прогнозирование в реальном времени**

* МЛ позволяет строить модели для оценки пористости и других параметров в реальном времени, что особенно ценно при бурении. Модели могут обучаться на данных со смежных скважин и давать моментальные прогнозы, улучшая управление разработкой месторождения [10].

6. **Снижение влияния человеческого фактора**

* МЛ-модели стандартизируют обработку данных и снижают риск ошибок, вызванных человеческим фактором, который может привести к погрешностям в расчетах. Таким образом, повышается надежность результатов [5].

7. **Оценка и устранение систематических ошибок**

* МЛ позволяет выявлять систематические отклонения и ошибки, присущие конкретным типам скважин или геологическим условиям. Например, если породы обладают специфическими свойствами, которые вызывают отклонения в показаниях каротажа, модель может обучиться корректировать эти особенности [7, 11].

Ограничения применения машинного обучения

* МЛ-модели нуждаются в большом количестве данных для обучения, особенно если данные очень разнообразны по геологическим условиям.
* МЛ подходит для прогнозирования в условиях, сходных с обучающими данными, но может дать неверные результаты при значительном изменении условий, поэтому требуется внимательный подход к выбору данных для обучения.

Применение машинного обучения для прогнозирования пористости и других параметров по ГИС может значительно повысить точность и надежность расчетов, особенно при интеграции с данными керна и учетах поправок на влияние глинистости и насыщенности.

## 2.2 Подготовка исходных данных для машинного обучения.

Целевой переменной, которая будет предсказываться на основании признаков (кривые ГИС) является коэффициент пористости, определенный на керне. Для получения набора данных (dataset) необходимо создать таблицу по всем скважинам с глубинами, на которых отбирался керн. На каждой глубине в таблицу сносятся признаки для обучения - кривые ГИС (АК, GK, NGK, BK, PL). Данные считывались с помощью библиотеки pandas в Python.

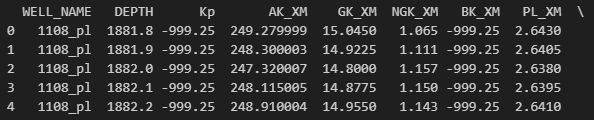


Рис. 2.2.1 Фрагмент загруженного dataset.

## 2.3. Разведочный анализ данных (EDA).

Разведочный анализ данных полученного dataset был выполнен для подготовки данных к обучению модели машинного обучения. Основные этапы EDA включали:

1. **Обзор данных**:
   * Оценка распределения данных, пропусков и выбросов.
   * Вычисление основных статистик (среднее, медиана, стандартное отклонение, минимальные и максимальные значения) [12, 13].
2. **Выявление выбросов**:
   * Использован метод межквартильного размаха (IQR) для поиска выбросов в признаках.
   * Выбросы заменены на ближайшие значения внутри допустимого диапазона.
3. **Корреляционный анализ**:
   * Вычислена корреляция признаков с целевой переменной. Признаки с низкой корреляцией (например, менее 0.1) проанализированы на предмет их полезности.
4. **Подготовка признаков**:
   * Для обучения модели были выбраны ключевые признаки: GK\_XM, NGK\_XM, BK\_XM, PL\_XM, AK\_XM.
5. **Визуализация**:
   * Построены графики распределения данных и взаимосвязей между признаками, а также графики ошибок модели и важности признаков после обучения.

EDA был выполнен несколькими итерациями. На первой итерации оценивались начальные загруженные данные для выявления их особенностей и понимания необходимости преобразования и дальнейшего использования в машинном обучении. После преобразования данных выполнялась вторая итерация для контроля результата и понимания необходимости дальнейшего преобразования.

На нижеследующем рисунке видно, что гистограмма значений целевой переменной содержит две основных корзины – с отрицательными значениями и со значениями, чуть более 1. Отрицательные значения – 999,25 соответствуют пропущенным значениям. Данные строки из dataset необходимо удалить.



Рис.2.3.1 Гистограмма значений целевой переменной Кп (Kp).

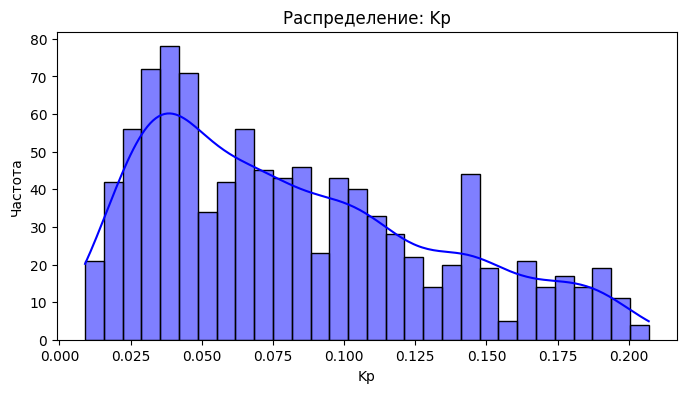


Рис.2.3.2 Гистограмма значений целевой переменной Кп (Kp) после удаления не существующих значений (-999,25).

В признаках тоже имелись несуществующие значения, т.к. для кривых ГИС свойственно иметь пропуски в разных скважина на разных глубинах ввиду различия методик исследований и брака полевого материала. Для признаков несуществующие значения не удален целыми строками из dataset, а преобразованы. Пропущенные значения признаков были заменены средними значениями. Стоит отметить, что каждый признак имел пропущенные значения. Ниже приведены гистограммы значений признака АК до преобразования пропущенных значений и после. Для остальных признаков показаны итоговые гистограммы.

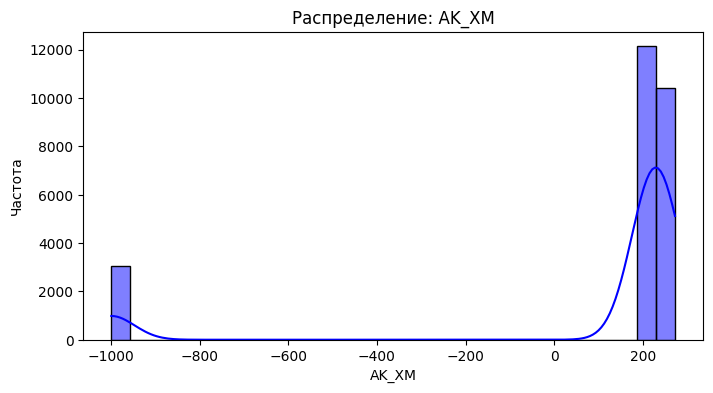


Рис.2.3.3 Гистограмма значений признака АК.

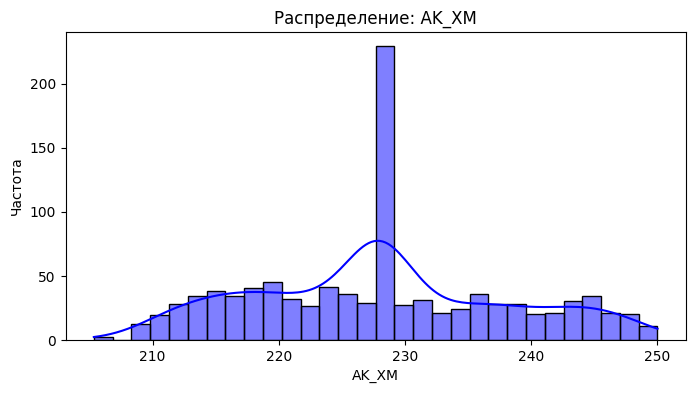


Рис.2.3.4 Гистограмма значений признака АК после преобразования пропущенных значений.

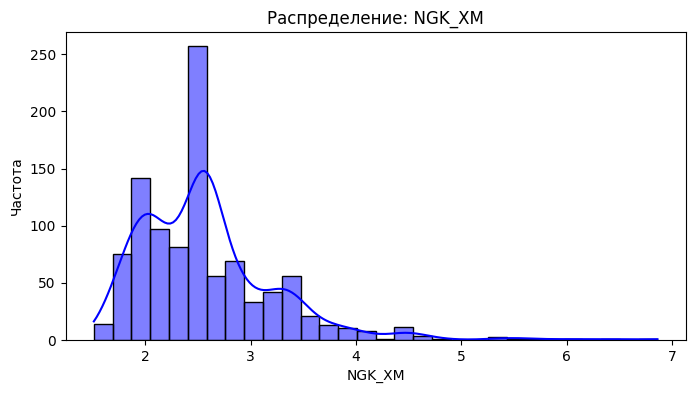


Рис.2.3.5 Гистограмма значений признака АК.

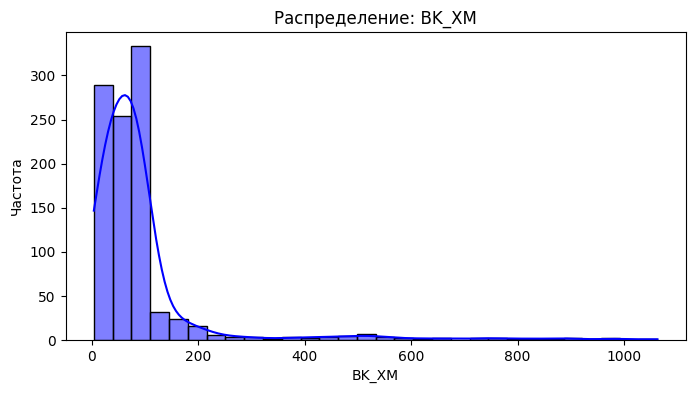


Рис.2.3.6 Гистограмма значений признака АК.

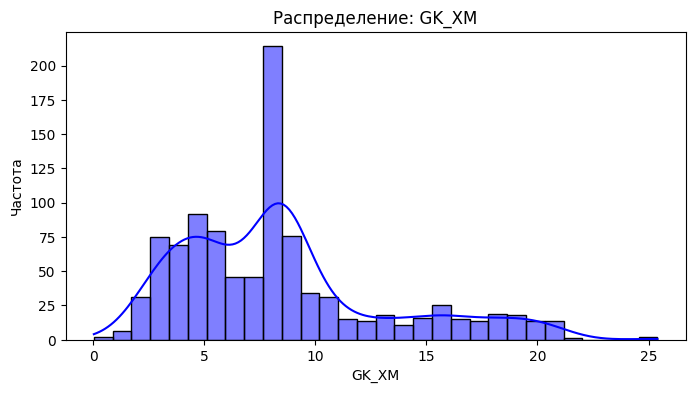


Рис.2.3.7 Гистограмма значений признака АК.

На 3 итерации EDA оценивались выбросы значений целевой переменной Кп и признаков для обучения. В значениях целевой переменной выбросов значений не было обнаружено.

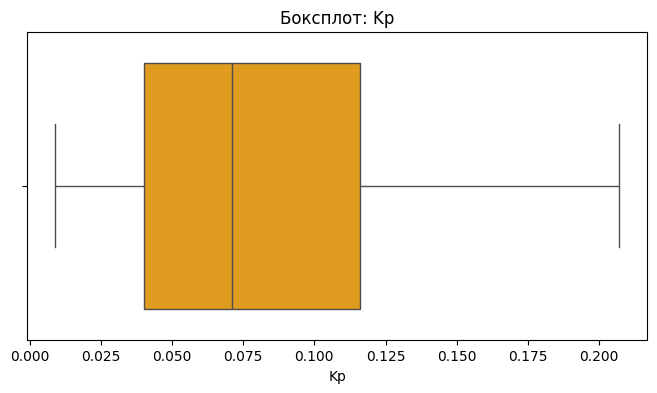


Рис.2.3.8 Боксплот для значений целевой переменной.

Для замены выбросов в признаках GK\_XM, NGK\_XM, BK\_XM, PL\_XM на ближайшие значения (границы диапазона без выбросов), был использован метод межквартильного размаха (IQR). Если значение выходит за пределы диапазона [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR], оно заменяется на ближайшую границу. Где Q1 – 25-й перцентиль, Q3 – 75-й перцентиль, IQR – межквартальный размах. Нижняя и верхняя границы вычисляются как (Q1 - 1.5 \* IQR) и (Q3 + 1.5 \* IQR) [14].

Для замены выбросов использовалась библиотека np.where [15] по следующим условиям:

- если значение меньше нижней границы, оно заменяется на lower\_bound.

- если значение больше верхней границы, оно заменяется на upper\_bound.

После выполнения кода все выбросы в указанных признаках были заменены на ближайшие границы диапазона. Нижи приведены рисунки для каждого признака до удаления выбросов и после.

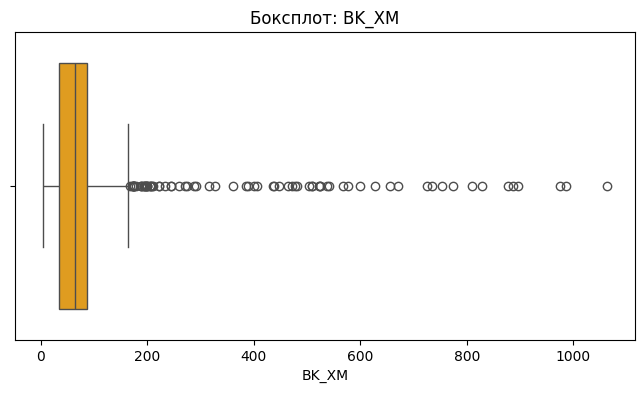


Рис.2.3.9 Боксплот для признака BK.

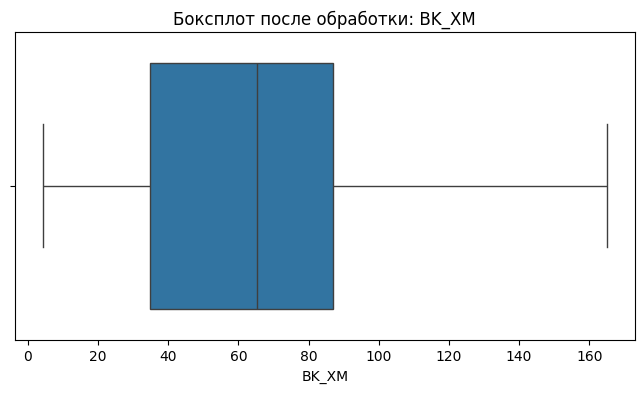


Рис.2.3.10 Боксплот для признака BK после замещения выбросов.

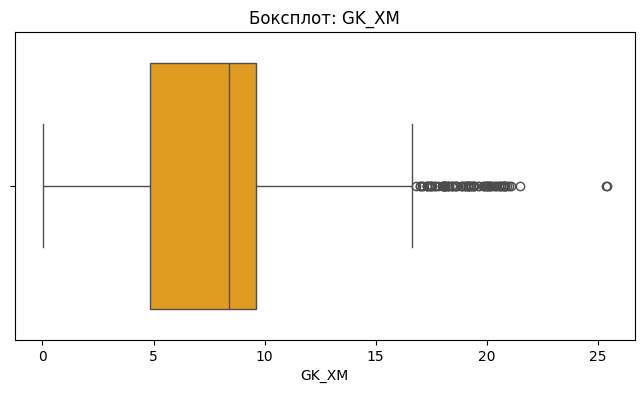


Рис.2.3.11 Боксплот для признака GK.

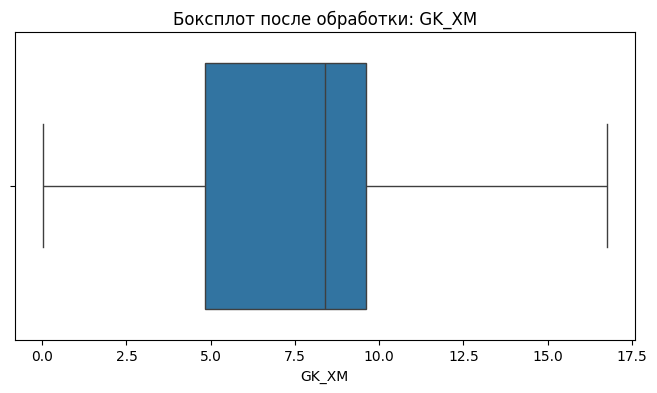


Рис.2.3.12 Боксплот для признака GK после замещения выбросов.

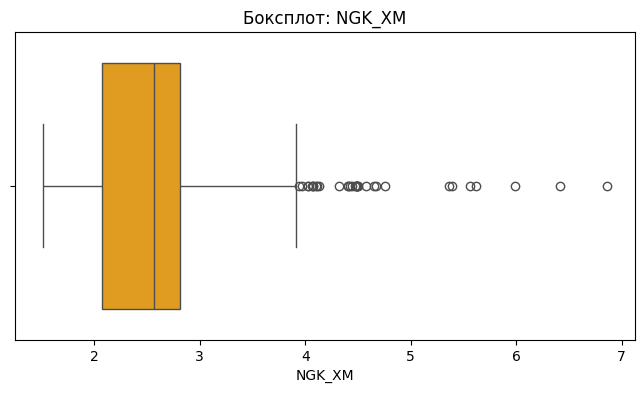


Рис.2.3.13 Боксплот для признака NGK.

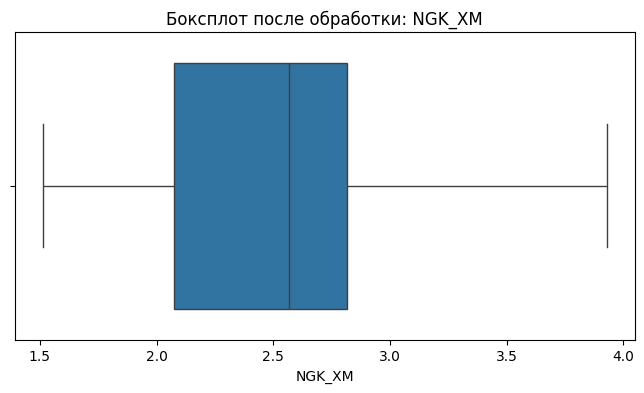


Рис.2.3.14 Боксплот для признака NGK после замещения выбросов.

Стоит отметить, что признак АК аномальных выбросов значений не имел (рисунок ниже).

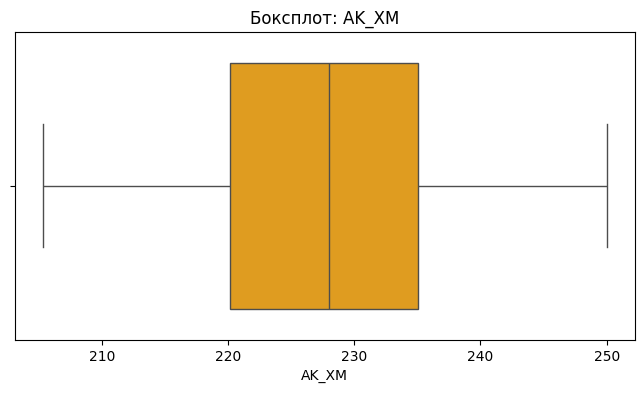


Рис.2.3.15 Боксплот для признака AK.

Особое внимание было уделено признаку плотностного каротажа – PL. Данный признак имеет большой разброс значений.

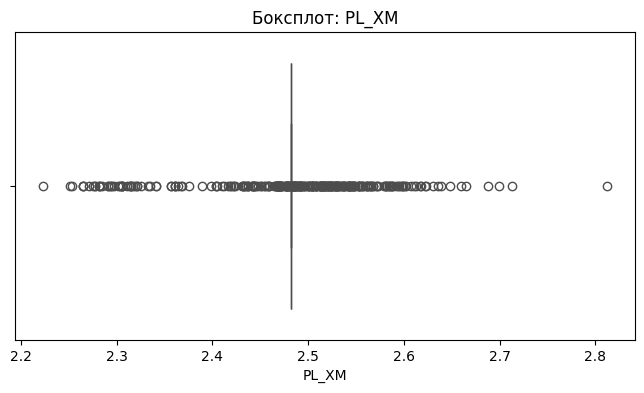


Рис.2.3.16 Боксплот для признака PL.

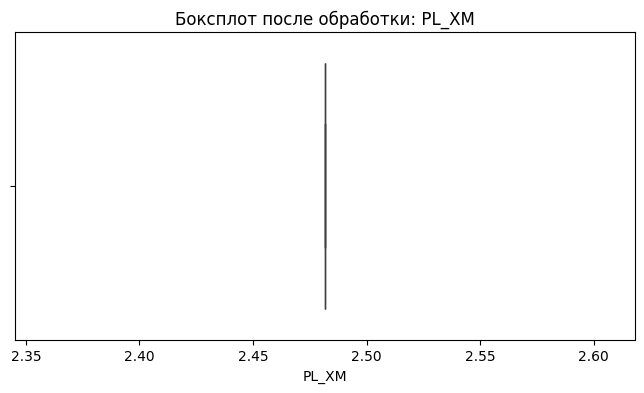


Рис.2.3.16 Боксплот для признака PL после замещения выбросов.

После устранения аномальных значений по той же технологии, что и для остальных признаков, осталось слишком мало значений. Проведена оценка статистических характеристик признака PL. Mean (стандартное отклонение) 2.481664, что указывает на низкую изменчивость данных., std 0.052799, min 2.223000, max 2.811970 узкий диапазон значений. Значения 25% 2.481664, 50% 2.481664, 75% 2.481664 совпадают, что говорит о высокой концентрации данных вокруг среднего значения.

Использование признака PL может быть проблематичным по следующим причинам:

- Низкая изменчивость (почти константное распределение). Такой признак может не вносить значимого вклада в обучение модели, так как он плохо дифференцирует данные.

- Большинство значений сосредоточено вокруг одного значения (2.481664), что снижает его информативность.

Когда признак может быть полезен только в случае если имеет смысловую интерпретацию, связанную с целевой переменной. Например, даже небольшие отклонения в значении могут быть важны для модели.

Критерии исключения признака из обучения следующие:

- Если анализ корреляции с целевой переменной показывает слабую или нулевую связь.

- Если использование этого признака не улучшает метрики модели или вносит шум.

Корреляция PL с целевой переменной: -0.100679. Значение указывает, что PL не имеет значительного линейного влияния на целевую переменную. Это может быть связано с несколькими причинами:

- Признак действительно мало влияет на целевую переменную.

- Связь может быть нелинейной, что не учитывается в метрике линейной корреляции.

Отрицательный знак eказывает, что увеличение значения PL\_XM слабо связано с уменьшением значения целевой переменной.

Таким образом на стадии EDA (3 итерация) принято решение оставить признак PL без замещения аномальными значениями до выяснение веса этого признака на результат обучения модели.

## 2.4. Выбор модели для обучения

Для выполнения задачи разбивки данных, обучения различных моделей, вычисления метрик и оценки значимости признаков, dataset был разбит на обучающую и тестовую выборки в соотношении 80:20. Далее было обучено несколько моделей (линейная регрессия , случайный лес, градиентный бустинг) [16, 17, 18]. Для оценки качества для каждой модели были рассчитаны метрики R², MAE, MSE.

Таблица 2.2.1. Результат оценки моделей.

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | R² | MAE | MSE |
| Linear Regression | 0.025873 | 0.039253 | 0.002252 |
| Random Forest | 0.527487 | 0.026535 | 0.001092 |
| Gradient Boosting | 0.387956 | 0.030904 | 0.001415 |

Для выбора модели проведен анализ. 1R² (Коэффициент детерминации) показывает, насколько хорошо модель объясняет изменчивость целевой переменной. Значение от 0 до 1, где 1 — идеальная модель. Для Linear Regression R2=0.0259, что указывает на почти полное отсутствие способности линейной регрессии предсказывать целевую переменную. Для Random Forest R2=0.5275, что демонстрирует умеренную способность объяснять целевую переменную. Для Gradient Boosting R2=0.3880, что хуже, чем у случайного леса, но лучше линейной регрессии.

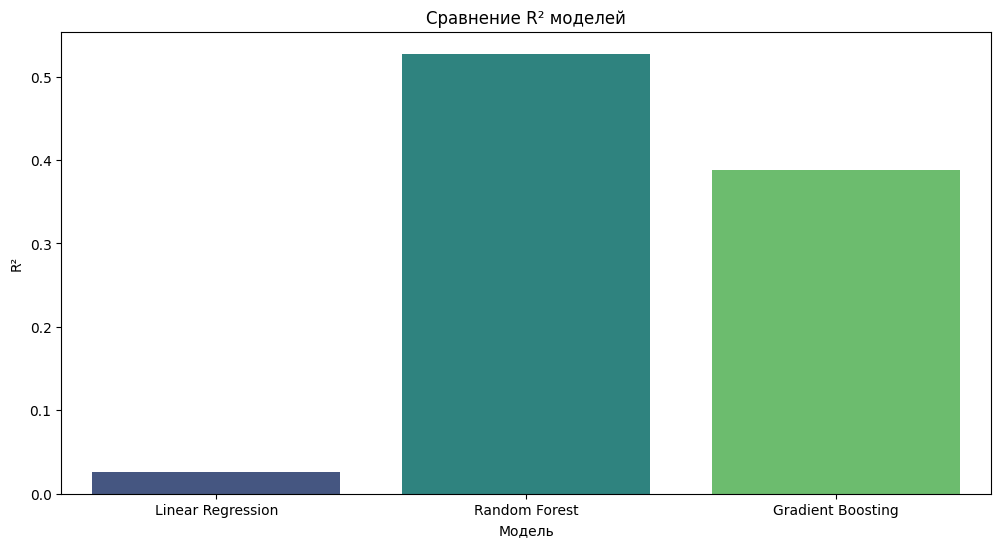


Рис.2.4.1 Сравнение коэффициента детерминации моделей.

MAE (Средняя абсолютная ошибка) поволяет оценить среднюю разницу между фактическими и предсказанными значениями. Чем меньше MAE, тем лучше модель. Random Forest имеет минимальное значение MAE (0.0265), что говорит о более точных прогнозах. Linear Regression и Gradient Boosting имеют более высокие значения MAE, что указывает на менее точные предсказания.

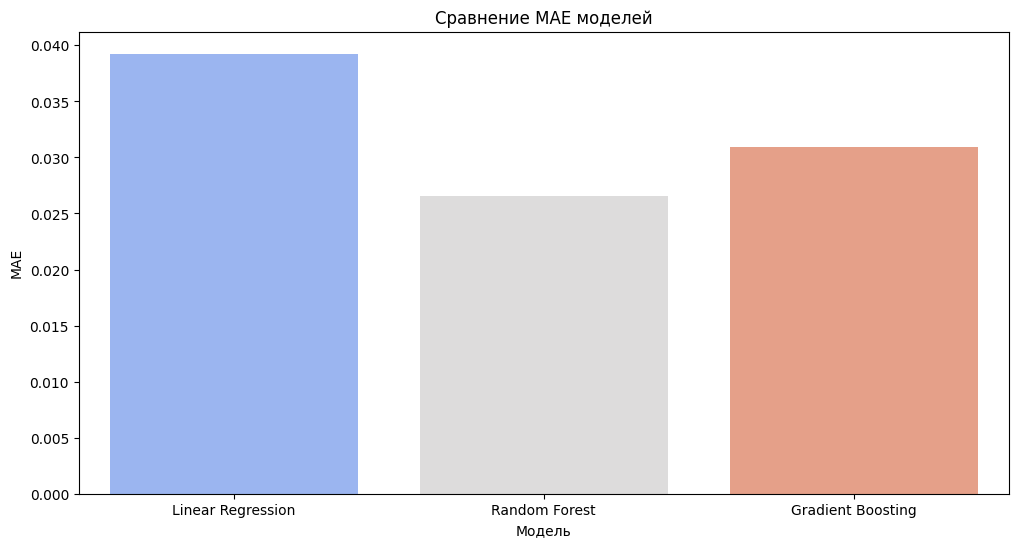


Рис.2.4.2 Сравнение метрик средней абсолютной ошибки моделей.

MSE (Среднеквадратичная ошибка) указывает на средний квадрат разницы между фактическими и предсказанными значениями. Меньшее значение MSE означает лучшую модель. Random Forest снова превосходит другие модели с наименьшим значением MSE (0.0011). Gradient Boosting имеет более высокую MSE (0.0014), но всё же лучше, чем у линейной регрессии.

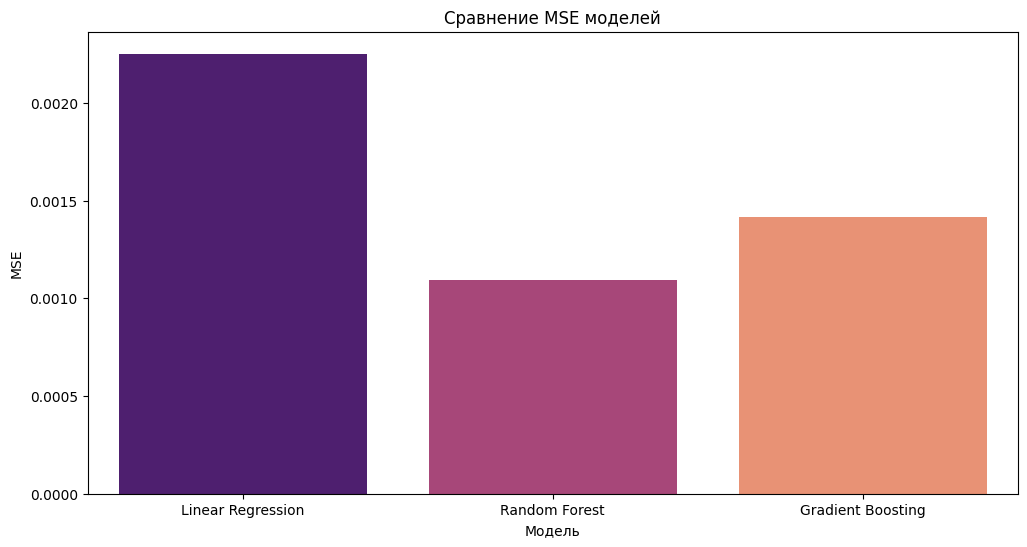


Рис.2.4.3 Сравнение метрик среднеквадратической ошибки моделей.

В результате проведенного анализа можно сделать следующие вывод:

Линейная регрессия практически не объясняет зависимость целевой переменной от признаков (R2≈0.03R²). Высокие MAE и MSE показывают низкую точность предсказаний. Линейная регрессия не рекомендуется для использования, так как она плохо работает с данной задачей.

Модель Random Forest показала лучшие результаты среди моделей по всем метрикам (R2=0.53R², MAE=0.0265MAE, MSE=0.0011MSE). Данная модель хорошо объясняет изменчивость целевой переменной, даёт точные предсказания. Модель рекомендуется для использования, особенно если нужна интерпретация важности признаков (позволит уточнить необходимость признака PL).

Модель Gradient Boosting показала второе место после случайного леса (R2=0.39R²). Хотя эта модель уступает случайному лесу по точности, она всё же значительно лучше линейной регрессии. Может быть полезной в задачах, где важна оптимизация гиперпараметров или снижение переобучения. Модель может быть использована, но её результаты хуже, чем у случайного леса.

Все пять признаков (GK\_XM, NGK\_XM, BK\_XM, PL\_XM, AK\_XM) оказались полезными, особенно в моделях, которые учитывают нелинейные зависимости (случайный лес и градиентный бустинг).

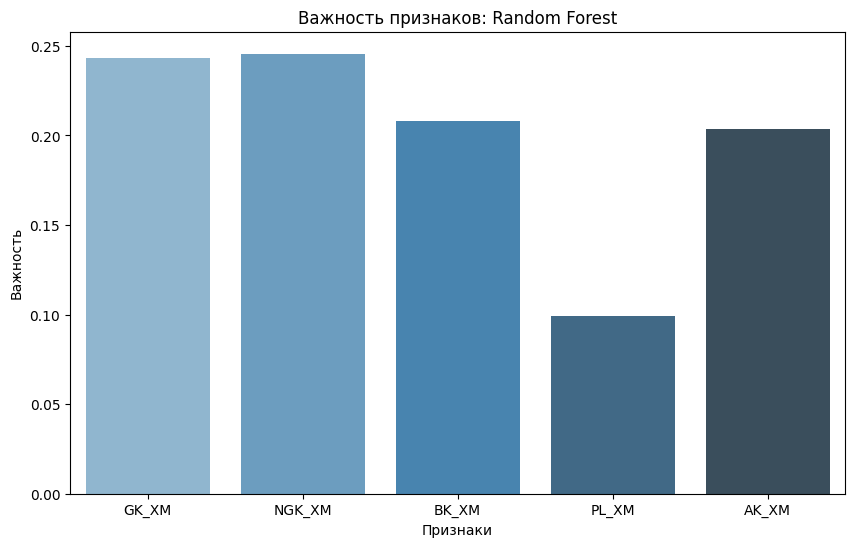


Рис.2.4.4 Важность признаков для модели Random Forest.

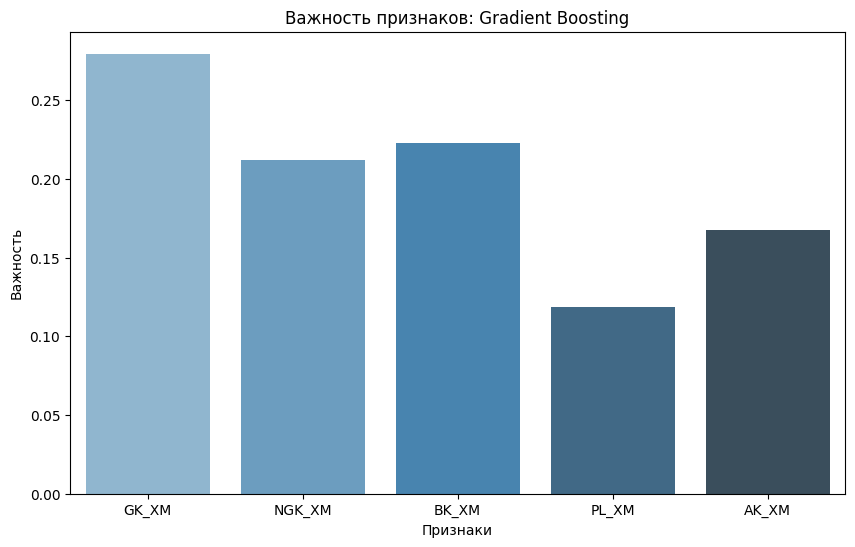
.

Рис.2.4.6 Важность признаков для модели Gradient Boosting.

Таким образом, для дальнейшего использования была выбрана модель Random Forest.

## 2.5. Подбор гиперпараметров для усовершенствования модели Random Forest.

Для подбора гиперпараметров модели Random Forest, можно использовать поиск по сетке (GridSearchCV) или случайный поиск (RandomizedSearchCV) из библиотеки sklearn. Для оптимизациивыбрано два гиперпараметра:

1. n\_estimators (количество деревьев).

2. max\_depth (глубина деревьев).

Для n\_estimators заданы колиства деревьев 50, 100, 200, 300 (чем больше деревьев, тем точнее, но медленнее обучение). Для max\_depth заданы None (без ограничения), 5, 10, 20, 30.

В результате получена модель со следующими лучшими параметрами max\_depth: None, n\_estimators: 300.

Max\_depth: None - модель не ограничивает глубину деревьев. Это значит, что деревья могут расти до полной глубины (до тех пор, пока каждый лист не станет чистым или не останется меньше заданного числа примеров).

n\_estimators: 300 - модель использует 300 деревьев. Это увеличивает точность, но также увеличивает время обучения.

Метрики лучшей модели следующие:

R²: 0.5410879532918251

MAE: 0.026171428893733445

MSE: 0.0010609411385620357

Значение R2 на тестовой выборке (20% от dataset) показывает, что модель объясняет 54.11% изменчивости целевой переменной. Это хороший результат для реальных данных.

Средняя абсолютная ошибка (MAE) составляет 0.02620, что указывает на низкую среднюю разницу между фактическими и предсказанными значениями.

Среднеквадратичная ошибка (MSE) также низкая, что подтверждает, что модель делает точные прогнозы.

Более наглядно результаты обучения модели представлены на нижеследующих рисунках. Для тестовой выборки изъято 20% данных.

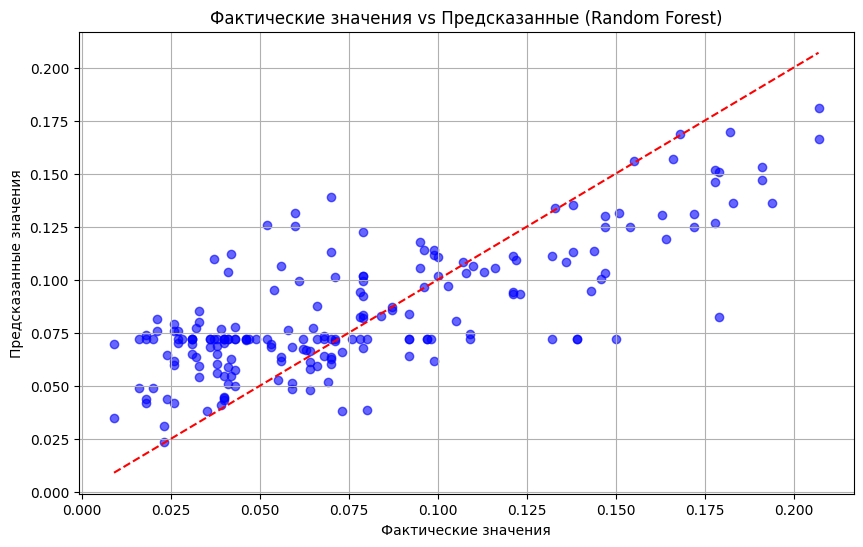
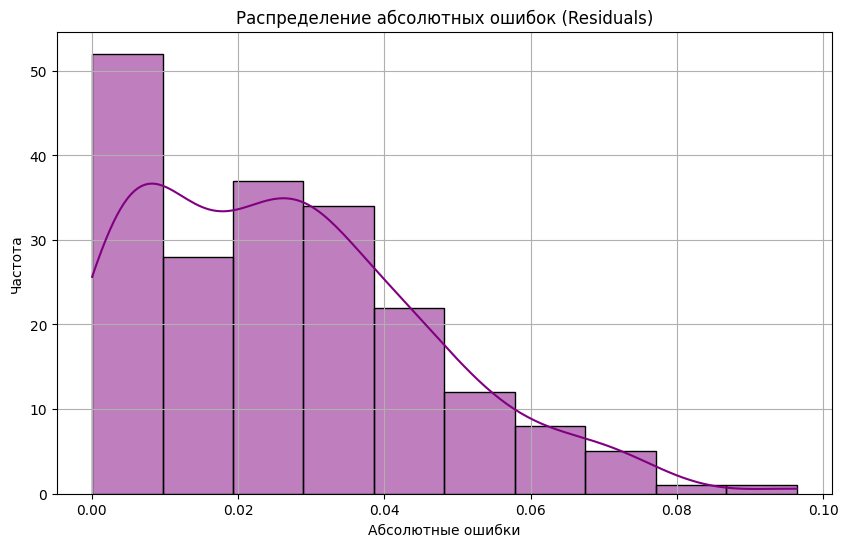


Рис.2.5.1 Сравнение фактических и предсказываемых значений. Тестовая выборка 20% от dataset.

 Рис.2.5.2 Гистограмма абсолютных ошибок.

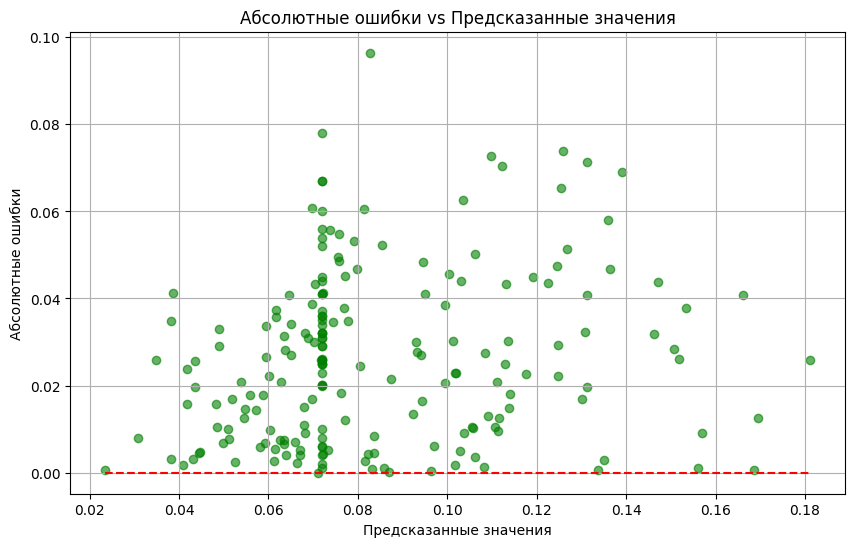


Рис.2.5.3 Сравнение абсолютных ошибок и предсказанных значений.

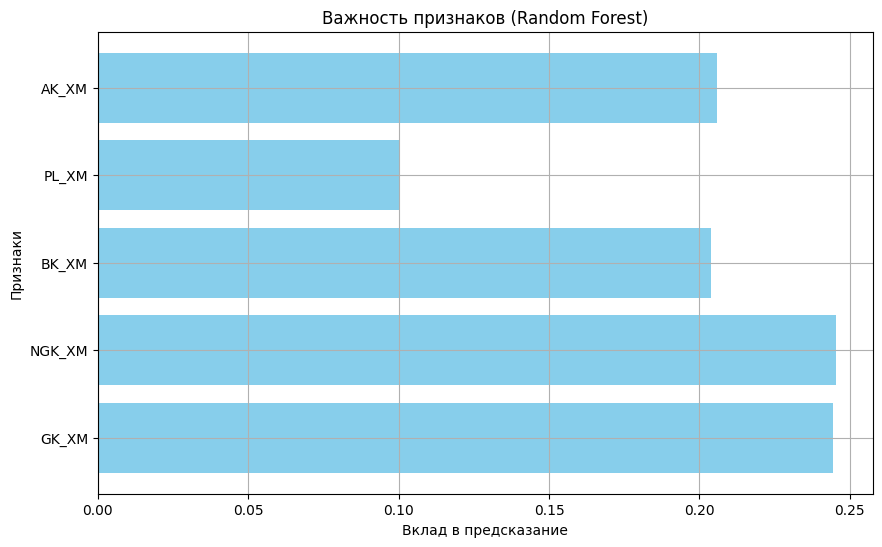


Рис.2.5.4 Важность признаков итоговой модели.

Таким образом, по итогу оценки качества модели можно сделать следующие вывод:

- Подобранные гиперпараметры привели к значительному улучшению метрик.

- Модель хорошо справляется с задачей, объясняя более 54% изменчивости целевой переменной.

- Низкие значения MAE и MSE делают её подходящей для практического применения.

- Признак PL наименее значимый в обучении. Проведены расчеты без его использования в обучении. Значительного улучшения не получено. Ввиду малого количества данных, способных обрабатываться незначительными вычислительными мощностями данный признак принято оставить.

## 2.6. Оценка модели на различных размерах тестовой выборки

Для оценки модели на различных размерах тестовой выборки (test\_size) проведено несколько экспериментов с разными пропорциями разделения данных на обучающую и тестовую выборки. На практике обычно используют следующие соотношения:

1. 80% обучающая / 20% тестовая (обычно используемое соотношение)

Это классический и часто используемый вариант. Он предоставляет достаточно данных для обучения модели и достаточно данных для проверки её обобщающей способности. Данный вариант использовался для подбора гиперпараметров итоговой модели для обучения.

2. 70% обучающая / 30% тестовая

Увеличенная доля тестовых помогает оценить модель на большем объеме тестовых данных. Для малого набора данных использование такого разбиения позволит оценить качество обобщения.

3. 90% обучающая / 10% тестовая

Это соотношение может быть полезно, когда для извлечения максимума из данных для обучения, но важно иметь небольшую, но достаточно репрезентативную тестовую выборку для оценки.

4. 50% обучающая / 50% тестовая

Это соотношение позволяет провести быструю оценку на малом наборе данных.

Был проведено тестирование модели с разными значениями test\_size, варьируя от 10% до 50%. Для каждой выборки выводятся метрики R², MAE, MSE, RMSE. Результаты приведены ниже на рисунке.

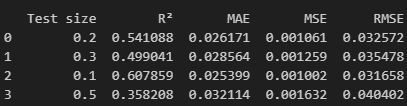


Рис. 2.6.1 Результаты тестирования модели с разными значениями test\_size.

Модель Random Forest имеет умеренно хорошее качество предсказаний для задачи с данным набором данных. Она объясняет около 54% вариации целевой переменной (по R2R^2R2) и имеет небольшие ошибки как по MAE, так и по MSE. Модель хорошо справляется с уменьшением ошибок (среднее отклонение около 0.026), что делает её достаточно стабильной. Высокая стабильность на всех метриках говорит о хорошем обучении модели на имеющихся данных.

## 2.7. Оценка применимости результатов моделирования в производстве.

Полученная по итогу обучения модель была использована для расчета коэффициента пористости для всего продуктивного интервала скважин. Результаты расчета загружены в постоянно действующую геолого-технологическую модель для тестирования на производственных задачах.

На нижеследующем рисунке показано сравнение коэффициентов корреляций коэффициентов пористости рассчитанных по кривым ГИС и по Кп по керну. К сожалению в конкретно в данном примере, Ккор полученный по стандартной методики расчета пористости слишком низкий – 0,17. Такой результат в работу брать нельзя, собственно по этой причине ввиду сжатых временных рамок было принято решение попробовать использовать машинное обучение для расчета коэффициентов пористости.

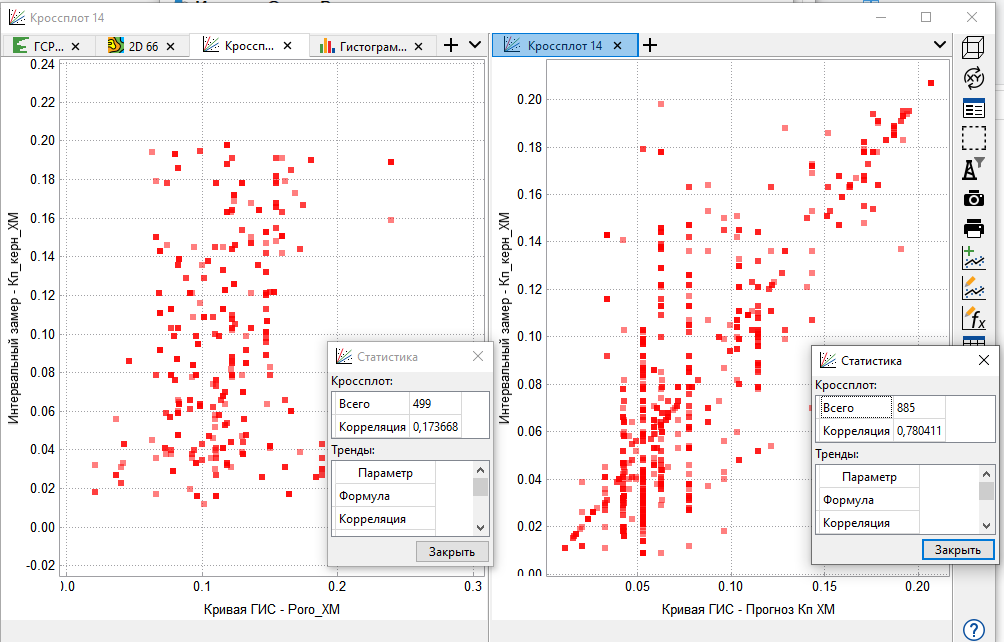


Рис. 2.7.1 Сравнение коэффициентов корреляций по Кп (стандартная методика) и Кп (машинное обучение).

На качественном уровне оценить результаты можно на рисунке 2.7.2. Хорошо видно, что Кп (машинное обучение) лучше повторяет значения Кп по керну, чем рассчитанные по стандартному подходу.

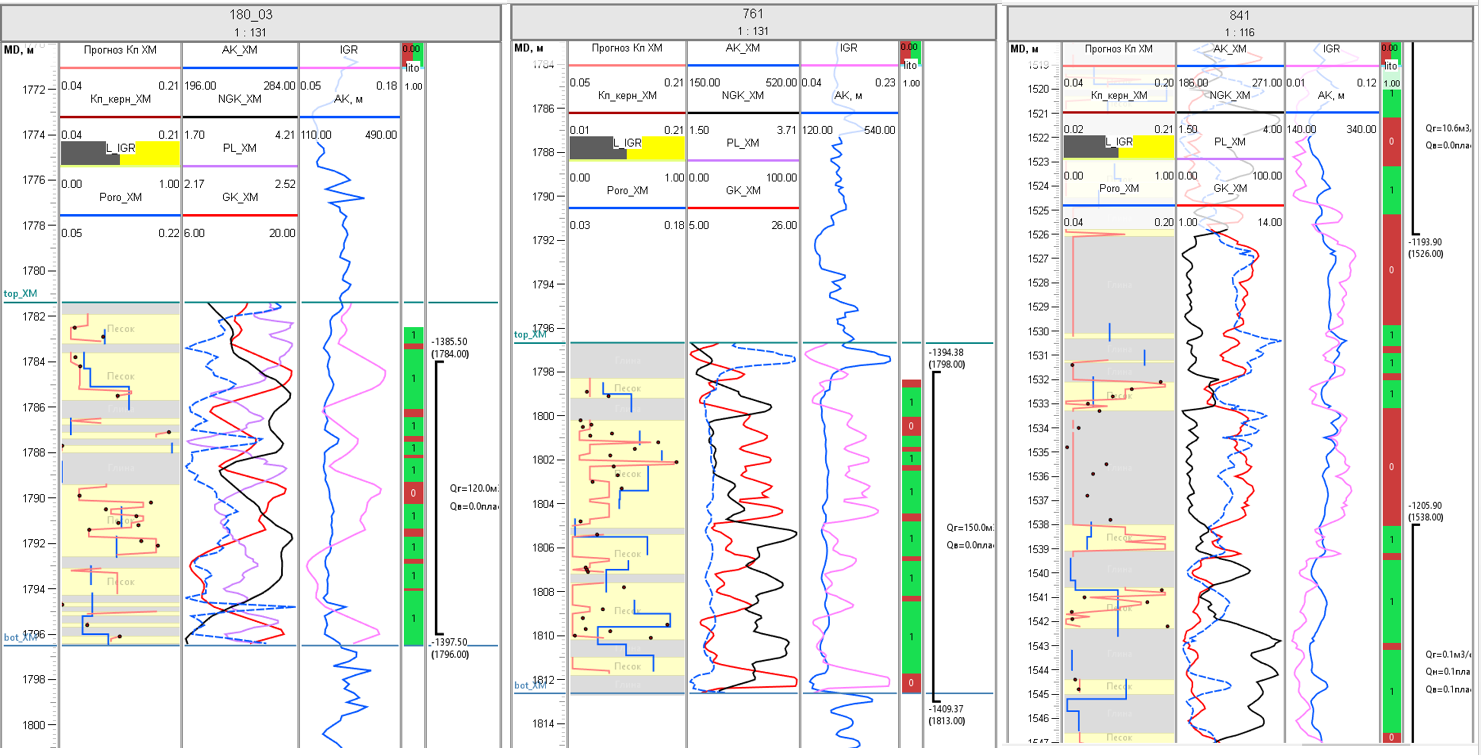


Рис. 2.7.2 Сравнение Кп керн (точки), Кп из принято ГМ (синяя линия), Кп машинное обучение (красная линия).

Дополнительно стоит отметить, что рассматриваемый пример месторождения углеводородов очень сложный по геологическому строению. Месторождение находится в активной стадии эксплуатационного разбуривания горизонтально направленными скважинами. Много скважин прошли по коллекторам с плохими ФЕС (низкими Кп). Полученная модель по итогу машинного обучения будет верифицироваться новыми скважинами.

Полученный расчет Кп использовался в двух производственных задачах – подсчет запасов углеводородов и проводке горизонтальных скважин.

Подсчет запасов – основной документ любого углевородододобывающего предприятия. От правильности подсчета запасов напрямую зависит экономический успех проекта. Зачастую геологи сильно ошибаются в запасах как в большую, так и в меньшую сторону. Есть месторождения на которых произведены капитальные затраты по строительству наземной инфраструктуры, заводов, трубопроводов, а запасов в итоге не хватает для загрузки мощностей. Есть обратная сторона – запасов по факту получается больше, инфраструктура не может обеспечить большую добычу. В итоге месторождение недозагружено. На одном из таких месторождений запасов хватает до 2600 г. Понятно, что целесообразность использования ископаемого топлива через несколько сотен лет стоит под большим вопросом. Использование машинного обучения дает прирост скорости в расчете Кп на несколько десятков порядков, что позволяет использовать различные варианты для оценки нефтегазового актива. Например, тот же Кп по крену определяется несколькими методами. До использования машинного обучения использование вариативности невозможно, ввиду отсутствия времени. В итоге, любое месторождение активно разбуривается на неповоротливых прогнозах. Отсутствует гибкость и, как следствие, корректировка экономических и технологических прогнозов добычи углеводородов.

Другая производственная задача, более мелкого масштаба – проводка горизонтальных скважин. Чтобы обеспечить запланированную проектом добычу по всему месторождению в целом, необходимо проложить горизонтальные скважины строго в эффективных породах. Грубо говоря – необходимом попасть скважиной в пласт. Чтобы попасть в пласт, нужно иметь цель. Цель в нашем привычном понимании в случае с месторождениями, которые находятся под землей, не видна прямыми методами наблюдений. Используются прогнозы целей под бурение. И вот возникают проблемы. Цели могут быть неверными как по площади, так и по разрезу. По площади – скважина может быть заложена в изначально низких запасах. По разрезу – скважина может пройти ниже пласта, не успеть набрать нужный угол, чтобы вернуться на цель и, как следствие, получить неэффективную проходку. Использование машинного обучения позволяет точнее определить цели с учетом геолого-геофизических неопределённостей.

В работе рассмотрен только коэффициент пористости. Не менее значимым параметрами является коэффициенты нефте- и газо-насыщенности.

## 2.8. Перспективы использования машинного обучения в нефтегазовой геологии.

Полученные результаты моделирования и накопленный опыт применения машинного обучения в производственных задачах открывают широкие возможности для дальнейшего развития и внедрения данных технологий в нефтегазовой отрасли. Основные направления использования машинного обучения включают:

* 1. Оценка неопределённости геолого-геофизических параметров при подсчете запасов на всех стадиях разведки и освоения месторождений. Появляется возможность оперативного расчета геологических моделей для оперативного анализа подтверждения прогнозов запасов в процессе получения новой геолого-геофизической информации.
  2. Создание альтернативных гидродинамических моделей для оценки всего нефтегазового актива и каждой скважины в отдельности. Например, использование проектных гидродинамических моделей для оценки дебита пробуренной скважины абсолютно неэффективно. Использование машинного обучения позволит учитывать весь набор имеющейся информации и получать более обоснованные прогнозы с возможностью их оперативной корректировки при накоплении информации.
  3. Геонавигационное сопровождение горизонтальны скважин с помощью машинного обучения позволит быстрее предсказывать изменение прогнозов, а значит своевременного реагировать на изменения, тем самым повышая эффективность проводки скважин по коллекторам с лучшими ФЕС.
  4. Инженерное и геомеханическое сопровождение скважин как вертикальных, так и горизонтальных, с использованием машинного обучения позволит выявить и предсказывать проблемы и осложнения бурения с акцентом на кривизну траекторий.
  5. Оптимизация разработки месторождений. Машинное обучение способствует разработке интегрированных моделей, объединяющих данные геологии, гидродинамики, бурения и эксплуатации скважин. Это позволяет комплексно подходить к оптимизации разработки, снижая затраты и увеличивая рентабельность проектов.
  6. Поддержка принятия решений. Интеграция машинного обучения в производственные процессы предоставляет возможность автоматизированного анализа больших объемов данных, выявления скрытых закономерностей и предложений оптимальных решений, основанных на моделях, обученных на исторических данных.

# Заключение

Пористость является основным параметром фильтрационно-емкостных свойств (ФЕС) горных пород залежей нефти и газа, который оказывает значительное влияние на объем запасов углеводородов и определяет экономическую эффективность разработки месторождений. Для повышения точности и скорости прогнозирования ФЕС необходимо внедрение современных технологий, таких как машинное обучение.

В данной работе обоснована целесообразность использования машинного обучения для расчета емкостных характеристик горных пород на основе нескольких каротажных методов, обязательную калибровку по керну. Этот подход позволяет существенно повысить надежность прогнозов.

Для достижения поставленной цели выполнены следующие задачи:

- подготовлен набор данных для машинного обучения;

- проведен разведочный анализ данных, что позволило выявить их статистические характеристики и определить применимость для построения моделей;

- выполнена обработка и подготовка признаков для обучения моделей;

- определены метрики для объективной оценки результатов прогнозирования;

- построены и проанализированы несколько моделей машинного обучения. Выбрана наиболее подходящая модель;

- усилена выбранная модель, проведена оценка прогноза и сделан вывод о ее применимости в производственном процессе.

Результаты работы подтверждают, что использование машинного обучения позволяет улучшить качество оценки пористости и других ФЕС горных пород, обеспечивая более точные расчеты и повышая эффективность разработки углеводородных месторождений.

Перспективное развитие технологий машинного обучения в нефтегазовой геологии создаёт платформу для повышения эффективности операций, минимизации рисков и улучшения управления ресурсами, что в конечном итоге способствует росту экономической выгоды проектов.

# Список литературы

* 1. Горшков А.А., Лапшин В.М. Методы интерпретации геофизических исследований скважин. — М.: Недра, 2005. — 350 с.
  2. Космовский А.М. Основы каротажных исследований: Учебное пособие для вузов. — СПб.: Лань, 2012. — 432 с.
  3. Майер В.Ф., Кузьмин В.А. Современные подходы к оценке пористости коллекторов по данным ГИС. — Уфа: Нефтегазовая геология, 2010. — 278 с.
  4. Роговская Н.А., Лопатин Ю.В. Интерпретация данных геофизических исследований скважин. — М.: Академкнига, 2016. — 295 с.

5. Aminzadeh, F., & Dasgupta, S. (2013). Geophysics for Petroleum Engineers. Society of Exploration Geophysicists.

6. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825–2830.

7. Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5–32.

8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

9. Kingma, D. P., & Ba, J. (2014). "Adam: A Method for Stochastic Optimization." arXiv:1412.6980 [cs.LG].

10. Chopra, S., et al. (2019). Advanced Data Analytics for Oil and Gas Exploration. Wiley.

11. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

12. McKinney, W. (2010). "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference, 51–56.

13. Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

14. SkyPro. (n.d.). Предобработка данных: очистка и нормализация. Retrieved January 14, 2025, from <https://sky.pro/wiki/python/predobrabotka-dannyh-ochistka-i-normalizaciya/?utm_source=chatgpt.com>

15. "Предобработка данных: очистка и нормализация." SkyPro. Web. 14 Jan. 2025. <https://sky.pro/wiki/python/predobrabotka-dannyh-ochistka-i-normalizaciya/?utm_source=chatgpt.com>.

16. Linear Regression: Scikit-learn. Модель LinearRegression. [Электронный ресурс]. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>.

17. Random Forest: Scikit-learn. Модель RandomForestRegressor. [Электронный ресурс]. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor>.

18. Gradient Boosting: Scikit-learn. Модель GradientBoostingRegressor. [Электронный ресурс]. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>.