Министерство науки и высшего образования 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

 «\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_г.

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**

**НАХОЖДЕНИЕ ОПТИМАЛЬНЫХ ГИПЕРПАРАМЕТРОВ ДЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДОБЫВАЮЩИХ СКВАЖИН**

По направлению подготовки

 01.03.03 «Механика и математическое моделирование»

Профиль

 01.03.03\_03 Математическое моделирование процессов нефтегазодобычи

Выполнил

Студент гр. 5030103/00301                                                            Кукуев А.И.

Руководитель

Старший преподаватель ВШТМиМФ                                        Симонов М.В.

Консультант

Главный специалист НОЦ «Газпромнефть-Политех»                 Печко К.А.

Санкт-Петербург

2024

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ   
УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО**

**Физико-механический институт**

**Высшая школа теоретической механики и математической физики**

УТВЕРЖДАЮ

Директор ВШТМиМФ

А. М. Кривцов

«\_\_» 20 г.

**ЗАДАНИЕ**

**на выполнение выпускной квалификационной работы**

студенту Кукуеву Артёму Игоревичу, гр. 5030103/00301

1. Тема работы: Нахождение оптимальных гиперпараметров для моделей машинного обучения добывающих скважин
2. Срок сдачи студентом законченной работы: 10.06.2024
3. Исходные данные по работе: актуальные научные публикации по теме работы, справочная литература, обезличенные реальные данные с нефтегазового месторождения
4. Содержание работы (перечень подлежащих разработке вопросов): разработка модели машинного обучения добывающих скважин на языке Python, сбор и анализ полученных гиперпараметров, формирование доверительных интервалов гиперпараметров
5. Перечень графического материала (с указанием обязательных чертежей): отсутствует
6. Консультанты по работе: Печко К.А., главный специалист НОЦ «Газпромнефть-Политех»
7. Дата выдачи задания 28.02.2024

Руководитель ВКР \_\_\_\_\_\_\_\_\_\_\_\_ Симонов М.В., старший преподаватель ВШТМиМФ

Задание принял к исполнению 28.02.2024

Студент \_\_\_\_\_\_\_\_\_\_\_\_ Кукуев А.И.

**РЕФЕРАТ**

На 41 стр., 18 рисунков, 3 таблицы.

МАШИННОЕ ОБУЧЕНИЕ, МЕТОДЫ ОПТИМИЗАЦИИ, ГИПЕРПАРАМЕТРЫ, PYTHON, ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Данная работа посвящена нахождению оптимальных гиперпараметров для моделей машинного обучения добывающих скважин. В рамках исследования рассматривались 4 метода оптимизации гиперпараметров моделей: Grid Search, Random Search, CmaEsSampler (CMA-ES) из библиотеки Optuna и Tree-structured Parzen Estimator (TPE) из библиотеки Hyperopt. Был проведен анализ данных с месторождений, и построены доверительные интервалы значений гиперпараметров для каждой из представленных моделей. Результаты представлены в виде графиков и таблиц, сравнивающих точность предсказанных значений целевой переменной и время, затраченное на работу каждого из методов.

**ABSTRACT**

41 pages, 18 pictures, 3 tables.

MACHINE LEARNING, OPTIMIZATION METHODS, HYPERPARAMETERS, PYTHON, CONFIDENCE INTERVALS

This work is devoted to finding optimal hyperparameters for machine learning models of producing wells. The study considered 4 methods for optimizing hyperparameters of models: Grid Search, Random Search, CmaEsSampler (CMA-ES) from the Optuna library and Tree-structured Parzen Estimator (TPE) from the Hyperopt library. The analysis of data from the deposits was carried out, and confidence intervals of hyperparameter values for each of the presented models were constructed. The results are presented in the form of graphs and tables comparing the accuracy of the predicted values of the target variable and the time spent on the operation of each of the methods.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc168440077)

[ГЛАВА 1. ОБЗОР СУЩЕСТВУЮЩИХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДОБЫВАЮЩИХ СКВАЖИН. 7](#_Toc168440078)

[ГЛАВА 2. ОБЩИЕ СВЕДЕНИЯ 10](#_Toc168440079)

[2.1. Процесс работы добывающей скважины 10](#_Toc168440080)

[2.2. Модель машинного обучения, параметры и гиперпараметры 14](#_Toc168440081)

[2.3. Описание методов оптимизации гиперпараметров 15](#_Toc168440082)

[2.4. Доверительный интервал 22](#_Toc168440083)

[2.5. Исходные данные с месторождений 24](#_Toc168440084)

[ГЛАВА 3. ПРАКТИЧЕСКАЯ ЧАСТЬ. 27](#_Toc168440085)

[3.1. Алгоритм программы для нахождения оптимальных гиперпараметров 27](#_Toc168440086)

[3.2. Результаты работы программы 28](#_Toc168440087)

[ЗАКЛЮЧЕНИЕ 38](#_Toc168440088)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 39](#_Toc168440089)

**ВВЕДЕНИЕ**

Успешная разработка и эксплуатация нефтяных и газовых месторождений в значительной степени зависит от эффективного применения передовых методов и технологий. В современных реалиях нефтегазовой отрасли все большее значение приобретают цифровые технологии, в частности методы машинного обучения. Они позволяют анализировать большие объемы данных, выявлять скрытые закономерности и строить прогнозные модели для различных производственных процессов.

Одной из ключевых задач, стоящих перед компаниями нефтегазовой отрасли, является оптимизация работы добывающих скважин. От эффективности и качества их эксплуатации напрямую зависят объемы добычи углеводородов и, как следствие, финансовые результаты деятельности предприятия. Традиционные методы анализа и прогнозирования, основанные на детерминированных моделях, зачастую не могут в полной мере учесть всю сложность и нелинейность процессов, происходящих в скважинах и пластах.

В связи с этим применение моделей машинного обучения для анализа данных о работе добывающих скважин представляется перспективным направлением. Однако эффективность таких моделей во многом определяется правильным подбором гиперпараметров – специальных параметров, настраиваемых вручную и влияющих на процесс обучения модели. От выбора оптимальных значений гиперпараметров зависят точность прогнозов, устойчивость модели к шумам в данных и ее способность к обобщению.

Целью данной дипломной работы является нахождение доверительных интервалов гиперпараметров для моделей машинного обучения добывающих скважин. Это позволит улучшить точность прогнозов, снизить риск переобучения модели и увеличить ее стабильность к изменениям входных данных.

Для достижения этой цели будут решены следующие задачи:

- Создание модели машинного обучения для добывающих скважин на языке Python, предсказывающей значение целевой переменной относительно параметров, регистрируемых телеметрией на месторождениях

- Проверка эффективности созданной модели на основе данных, полученных с разных месторождений

- Сбор и проведение анализа гиперпараметров, используемых в модели, и формирование выборки их значений

- Проведение статического анализа и построение доверительных интервалов для гиперпараметров модели

**ГЛАВА 1. ОБЗОР СУЩЕСТВУЮЩИХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДОБЫВАЮЩИХ СКВАЖИН.**

В настоящее время, с развитием технологий искусственного интеллекта и машинного обучения, все большее число предприятий нефтегазовой отрасли обращают внимание на применение этих технологий для оптимизации процессов добычи углеводородов. Одним из направлений применения машинного обучения в нефтегазовой отрасли является разработка моделей прогнозирования различных параметров скважин, таких как дебит, давление, температура и т. д.

Целью данного литературного обзора является изучение существующих моделей, оптимизирующих процессы нефтегазодобычи.

Для начала, рассмотрим работу [12], в которой авторы использовали глубокое обучение, как один из способов прогнозирования забойного давления при вертикальном многофазном потоке. Они провели большой объем работы по подбору структуры полносвязной нейронной сети и сравнили ее с классическими гидравлическими корреляциями. В итоге модель показала наибольшую способность прогнозирования по сравнению с классическими инженерными подходами.

В другой статье [9] рассмотрено использование различных способов и подходов для поиска забойного давления для газодобывающих скважин. В этой работе были проведены различные эксперименты, рассмотрены как стандартные модели машинного обучения, в частности, линейные регрессии типа Lasso и Ridge, так и методики глубокого обучения. В результате исследований они пришли к выводу, что лучшей из тестируемых моделей была нейронная сеть.

В 2018 году группа российских исследователей опубликовали работу [11], в которой описали процесс разработки прогнозирующей модели для многофазных потоков в стволе скважины, основанной на наборах деревьев решений. Они использовали данный комплекс для нахождения одного из ключевых параметров потока в стволе скважины, а именно забойного давления. По итогам проведенных исследований был сделан вывод о том, что их модель выдает достаточно хорошие показатели нормализованной среднеквадратичной ошибки (NRMSE) по сравнению с использованием нейронных сетей. Более того, полученное ими решение является более масштабируемым и демонстрирует хорошую шумоустойчивость.

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Рисунок 1.1. Среднее (слева) и наихудшее (справа) предсказанные тестовые выборки для случайного леса в работе [11]

Рассмотрим использование моделей машинного обучения для прогнозирования динамического пластового давления на примере работы [8]. В данной статье авторы проводят исследования, направленные на разработку косвенной методики определения пластового давления без остановки скважин. В качестве математической основы они используют два метода искусственного интеллекта – многомерный регрессионный анализ и нейронная сеть. Комбинирование данных методик демонстрирует достаточную работоспособность и позволяет достоверно определять пластовое давление даже при минимальном наборе исходных данных.

Изображение выглядит как линия, График, скат, диаграмма

Автоматически созданное описание

Рисунок 1.2. Сопоставление динамики фактического и модельного пластовых давлений за всю историю эксплуатации скважины в работе [8]

В связи с рассмотренными выше работами можно сделать вывод, что использование моделей машинного обучения для оптимизации процессов нефтегазодобычи является перспективным направлением. Однако, для того чтобы модели могли эффективно работать, необходимо тщательно подбирать гиперпараметры, которые влияют на их точность и скорость работы.

Исследование, проведенное в рамках данной работы, позволит не только оптимизировать работу существующих моделей машинного обучения в нефтегазовой отрасли, но и разработать новые, более эффективные модели, которые будут способны решать более сложные задачи. Это, в свою очередь, должно привести к увеличению добычи углеводородов и оптимизации работы в целом.

**ГЛАВА 2. ОБЩИЕ СВЕДЕНИЯ**

* 1. **Процесс работы добывающей скважины**

Процесс работы добывающей скважины направлен на извлечение углеводородов из недр земли. При этом, для обеспечения эффективной и безопасной добычи необходимо контролировать и оптимизировать различные параметры процесса, такие как забойное давление, устьевое давление, дебит жидкости, газовый фактор и обводненность [1].

Забойное давление — давление флюида на забое эксплуатируемой нефтяной, газовой или водяной скважины; характеризует энергию пласта, обусловливающую подъём жидкости (или газа) в стволе скважины. Рассчитывается по формуле (для неподвижного столба однокомпонентной жидкости):

где

– давление на забое скважины, Па;

– высота столба жидкости в скважине, м;

– плотность жидкости, кг/м3;

*–* давление на устье скважины, Па.

В формуле расчета забойного давления (1) для неподвижного столба однокомпонентной жидкости потери на трение не учитываются. Однако, в реальных условиях добычи углеводородов, где жидкость и газ движутся в стволе скважины, потери на трение являются существенным фактором, влияющим на забойное давление.

Для учета потерь на трение при движении флюида в скважине используется формула Дарси-Вейсбаха. Эта формула была получена эмпирическим путем в 19-м веке немецким инженером Юлиусом Вейсбахом, который проводил эксперименты по изучению турбулентного течения воды в трубах. Вейсбах обнаружил, что потери на трение при турбулентном течении пропорциональны квадрату скорости флюида и обратно пропорциональны диаметру трубы. Он выразил эти зависимости в виде формулы, которая позже была названа его именем. Рассмотрим данную формулу в дифференциальном виде:

где

– градиент давления, Па/м;

– коэффициент трения, зависящий от режима течения и шероховатости стенки скважины;

–плотность флюида, кг/м3;

– скорость флюида, м/с;

– диаметр скважины, м.

В более сложных моделях скважин, которые учитывают многофазное течение, взаимодействие фаз, турбулентность и другие факторы, используются более сложные дифференциальные уравнения и численные методы их решения.

Таким образом, потери на трение являются важным фактором, влияющим на забойное давление и эффективность добычи углеводородов.

Устьевое давление — давление в верхней точке скважины, на ее устье; измеряется манометрами устьевой арматуры. Различают статические и динамические устьевые давления.

Статическое устьевое давление замеряется в остановленной скважине и зависит от пластового давления, глубины скважины и плотности заполняющей ее среды. Оно численно равно разности пластового давления и давления столба жидкости от устья до пласта.

Динамическое устьевое давление измеряется в действующей скважине, зависит от тех же параметров, что и статическое, и, кроме того, от дебита скважины или расхода нагнетательного агента, а также от давления в трубопроводе у скважины и перепада давлений в запорно-регулирующих органах устьевой арматуры. Избыточное устьевое давление по отношению к атмосферному может достигать 100 МПа и более (в газовых скважинах, при гидроразрыве пласта).

Изображение выглядит как дизайн

Автоматически созданное описание

Рисунок 1.1. Упрощенная схема скважины

Дебит — это количество жидкости или газа, которое извлекается из скважины за единицу времени. Этот параметр зависит от многих факторов, таких как свойства пласта, свойства жидкости и газа, забойное давление, устьевое давление, и т. д. Контроль и оптимизация дебита необходимы для обеспечения эффективной добычи углеводородов. Для расчета дебита нефтяной скважины используется стандартная формула:

где

– дебит;

– высота столба жидкости в скважине, м;

– производительность насоса;

*–* статический уровень, расстояние от начала подземных вод до первых слоёв почвы;

*–* динамический уровень, абсолютная величина, получаемая при замере уровня воды после откачивания.

Газовый фактор — это отношение объема газа, растворенного в жидкости, к объему этой жидкости. Этот параметр важен при контроле и оптимизации процесса добычи, так как влияет на эффективность транспортировки углеводородов из скважины на поверхность. Для нефтяных месторождений нашей страны газовый фактор изменяется от 20 до 1000 м3/т (в среднем он составляет 100 м3/т). Этот параметр обычно определяется по пробам пластовой нефти путем её дегазации при нормальной температуре.

Обводненность скважины — это содержание воды в продукции скважины, определяемое как отношение дебита воды к сумме дебитов нефти и воды. Обводненность скважин определяют систематическим отбором проб жидкости, поступающей из скважин, и автоматическим контролем за обводненностью.

Характер обводнения пластов-коллекторов различен - он зависит от свойств продуктивных пластов, начальных условий залегания нефти в пласте и системы разработки нефтяных месторождений.

* 1. **Модель машинного обучения, параметры и гиперпараметры**

В этой главе мы рассмотрим модель машинного обучения, которую мы будем использовать для прогнозирования забойного давления в добывающих скважинах. Кроме того, мы обсудим гиперпараметры, которые являются ключевыми компонентами любой модели машинного обучения, и которые мы будем оптимизировать в нашей работе.

Модель машинного обучения представляет собой алгоритм, который использует данные для обучения и предсказания результатов на основе этих данных [5]. В нашей работе мы будем использовать модель градиентного бустинга от sklearn (Gradient Boosting Regressor, GBR), которая является одной из самых точных и эффективных моделей машинного обучения.

GBR является моделью ансамблевого обучения, которая объединяет множество слабых моделей для создания одной сильной модели. Слабые модели в GBR представляют собой регрессионные деревья, которые строятся последовательно, с учетом ошибок, сделанных предыдущими деревьями. Таким образом, каждое новое дерево улучшает предсказание модели.

Изображение выглядит как снимок экрана, диаграмма, линия, Шрифт

Автоматически созданное описание

Рисунок 1.2. Визуальная интерпретация метода градиентного бустинга.

Гиперпараметры и параметры — это два ключевых аспекта, которые формируют модель машинного обучения. Они играют разные роли в процессе обучения и влияют на поведение и производительность модели.

Гиперпараметры — это настройки модели, которые определяют её общую структуру и способ обучения [4]. Эти параметры устанавливаются до начала процесса обучения и не изменяются в процессе обучения модели. Они оказывают влияние на то, как модель будет обучаться, какие признаки будут учитываться, и какие ограничения будут наложены на процесс обучения.

Параметры модели, с другой стороны, являются внутренними весами или коэффициентами, которые модель обучает в процессе обучения на основе обучающих данных. Эти параметры изменяются в процессе обучения с целью минимизации функции потерь и достижения наилучшего соответствия между прогнозами модели и реальными значениями целевой переменной.

Различие между гиперпараметрами и параметрами модели заключается в том, что первые задаются вручную до начала обучения и определяют характеристики всего процесса обучения, в то время как параметры модели вычисляются в процессе обучения на основе данных и оптимизируются для достижения наилучшей производительности. Гиперпараметры можно сравнить с настройками инструмента, с помощью которого вы создаете модель, а параметры модели — это результат работы этого инструмента на конкретных данных.

* 1. **Описание методов оптимизации гиперпараметров**

В рамках задачи оптимизации гиперпараметров модели были использованы следующие методы:

* Grid Search
* Random Search
* CmaEsSampler из библиотеки Optuna
* TPE из библиотеки Hyperopt

Grid Search (сетчатый поиск) является одним из наиболее простых и распространенных методов оптимизации гиперпараметров. Суть метода заключается в том, что задается сетка значений гиперпараметров, и для каждой комбинации значений проводится обучение модели. После обучения всех моделей выбирается та, которая показала наилучшие результаты на валидационном наборе данных.

Random Search (случайный поиск) представляет собой модификацию метода Grid Search. В отличие от Grid Search, значения гиперпараметров выбираются случайным образом из заданного диапазона. После обучения всех моделей выбирается та, которая показала наилучшие результаты на валидационном наборе данных.

CmaEsSampler является одним из методов оптимизации гиперпараметров, доступных в библиотеке Optuna. Он использует эволюционный алгоритм для поиска оптимальных гиперпараметров.

Эволюционный алгоритм, используемый в CmaEsSampler, называется Covariance Matrix Adaptation Evolution Strategy (CMA-ES). Он работает путем создания популяции кандидатов в гиперпараметры, затем оценивая их качество с помощью функции целевого параметра, и затем использует информацию о качестве для выбора новых кандидатов в гиперпараметры.

Формально, пусть - вектор гиперпараметров, - функция, которая оценивает качество модели при заданных гиперпараметрах, и D - набор данных, используемых для обучения модели. CMA-ES строит популяцию кандидатов в гиперпараметры, затем оценивает их качество с помощью , и затем использует информацию о качестве для выбора новых кандидатов в гиперпараметры.

Одной из ключевых особенностей CMA-ES является адаптация ковариационной матрицы. Ковариационная матрица описывает связь между гиперпараметрами, и адаптация ковариационной матрицы позволяет CMA-ES эффективно искать оптимальные гиперпараметры в пространствах с высокой размерностью.

Другой важной особенностью CMA-ES является адаптивный шаг обучения. Вместо того, чтобы использовать фиксированный шаг обучения для всех гиперпараметров, CMA-ES адаптирует шаг обучения для каждого гиперпараметра в зависимости от его вклада в общий результат. Это позволяет CMA-ES сходиться к оптимальным гиперпараметрам быстрее, чем другие методы оптимизации гиперпараметров.

Описание алгоритма CMA-ES можно представить в следующем виде:

1. Инициализация:

Создается начальная популяция кандидатов в гиперпараметры , где n - размерность популяции. Задается начальная ковариационная матрица и начальный вектор шага .

1. Оценка качества:

Для каждого кандидата в гиперпараметры вычисляется значение целевой функции , которая оценивает качество модели при заданных гиперпараметрах.

1. Отбор и обновление ковариационной матрицы:

Выбирается подмножество из наилучших кандидатов в гиперпараметры , где m - размерность подмножества. Затем вычисляется новая ковариационная матрица с помощью следующей формулы:

где

- коэффициент, отвечающий за скорость обновления ковариационной матрицы,

- вероятность выбора шага, - среднее значение наилучших кандидатов в гиперпараметры.

1. Обновление вектора шага:

Вычисляется новое значение вектора шага с помощью следующей формулы:

где

и - коэффициенты, отвечающие за скорость обновления вектора шага, - норма вектора ,

- математическое ожидание нормы вектора .

1. Обновление популяции:

Создается новая популяция кандидатов в гиперпараметры , с помощью следующей формулы:

где

- нормальное распределение с нулевым средним и ковариационной матрицей .

1. Проверка условия остановки:

Если выполнено условие остановки (например, достигнуто максимальное количество итераций или найден достаточно хороший кандидат в гиперпараметры), то алгоритм завершается. Иначе, переходим к шагу 2.

Tree-structured Parzen Estimator (TPE) из библиотеки Hyperopt — это метод оптимизации гиперпараметров, основанный на байесовском подходе. Он был разработан для решения задач оптимизации в пространствах с большим числом гиперпараметров. Алгоритм строит модель распределения гиперпараметров, которая максимизирует функцию целевого параметра, и использует ее для предсказания наилучших значений гиперпараметров. После обучения модели с новыми значениями гиперпараметров модель распределения обновляется, и процесс повторяется.

TPE использует два распределения для моделирования функции целевого параметра от гиперпараметров: одно распределение для хороших значений гиперпараметров (l) и одно для плохих значений гиперпараметров (g). Каждое распределение представляет собой смесь мультивариатных нормальных распределений, которые моделируют подмножества хороших и плохих гиперпараметров.

Формально, пусть x - вектор гиперпараметров, y - значение целевого параметра, - множество хороших гиперпараметров, - множество плохих гиперпараметров. Требуется найти такой вектор x, который минимизирует y.

TPE использует следующую модель для оценки вероятности того, что гиперпараметры x принадлежат множеству хороших гиперпараметров :

где

– i-й гиперпараметр

- условная вероятность того, что i-й гиперпараметр принадлежит множеству хороших гиперпараметров, при условии того, что известны значения предыдущих гиперпараметров.

Аналогично, TPE использует следующую модель для оценки вероятности того, что гиперпараметры x принадлежат множеству плохих гиперпараметров :

TPE использует эти модели для оценки вероятности того, что гиперпараметры x приведут к хорошему или плохому значению целевого параметра y. Затем, TPE использует эту информацию для выбора нового вектора гиперпараметров x, который, как ожидается, приведет к улучшению значения целевого параметра.

TPE использует два типа стратегий для выбора нового вектора гиперпараметров: эксплуатацию (exploitation) и исследование (exploration). Эксплуатация состоит в выборе гиперпараметров, которые, как ожидается, приведут к хорошему значению целевого параметра, в соответствии с текущей моделью. Исследование состоит в выборе гиперпараметров, которые, как ожидается, улучшат текущую модель, даже если они могут привести к плохому значению целевого параметра.

TPE использует следующую формулу для выбора нового вектора гиперпараметров x:

где

p(x) - вероятность выбора гиперпараметров x

α - параметр, который контролирует баланс между эксплуатацией и исследованием.

Рассмотрим таблицу 2.1, в которой представлено сопоставление методов оптимизации и описаны преимущества и недостатки каждого из них.

Таблица 2.1. Сравнение методов оптимизации гиперпараметров

|  |  |  |
| --- | --- | --- |
| Метод оптимизации | Преимущества | Недостатки |
| Grid Search | - Простота реализации  - Высокая точность результатов | - Высокая вычислительная сложность  - Неэффективен при большом количестве гиперпараметров |
| Random Search | - Меньшая вычислительная сложность по сравнению с Grid Search  - Может найти хорошие гиперпараметры при большом количестве параметров | - Не гарантирует нахождение оптимальных гиперпараметров  - Может потребоваться больше времени для поиска хороших гиперпараметров |
| CmaEsSampler (CMA-ES) из библиотеки Optuna | - Наличие встроенных визуализаций  - Эффективен при большом количестве гиперпараметров  - Имеет поддержку распределенных вычислений | - Относительно высокие требования к ресурсам  - Относительно сложный синтаксис и недостаток документации  -Требует увеличенного количества времени на настройку |
| Tree-structured Parzen Estimator (TPE) из библиотеки Hyperopt | - Использует алгоритм деревьев решений  - Имеет возможность сохранять и загружать результаты оптимизации  - Позволяет задавать условия для поиска гиперпараметров | - Более сложен в реализации по сравнению с Grid Search и Random Search  - Отсутствие встроенных визуализаций |

* 1. **Доверительный интервал**

В этой главе мы сосредоточимся на оценке качества найденных оптимальных гиперпараметров с помощью доверительного интервала.

Доверительный интервал (Confidence Interval, CI) — это статистический интервал, который используется для оценки надежности оценки параметра [6]. Его будем использовать для оценки надежности оптимальных гиперпараметров, найденных с помощью различных методов оптимизации. Он строится на основе выборки и характеризуется двумя параметрами: уровнем доверия (confidence level) и шириной интервала (interval width). Уровень доверия определяет вероятность того, что истинное значение параметра попадает в доверительный интервал. Ширина характеризует размер интервала, в котором, согласно уровню доверия, находится истинное значение параметра.

Для построения доверительного интервала необходимо выбрать соответствующую статистическую модель и определить ее параметры. Мы будем использовать нормальную модель, которая широко применяется в статистическом анализе. Нормальная модель характеризуется двумя параметрами: математическим ожиданием (средним значением) и стандартным отклонением.

Для оценки оптимальных гиперпараметров мы будем использовать среднее значение, а для построения доверительного интервала - стандартное отклонение. Стандартное отклонение характеризует степень рассеивания выборки вокруг среднего значения. Чем меньше стандартное отклонение, тем выше точность оценки параметра.

Для построения доверительного интервала с уровнем доверия α мы будем использовать следующую формулу:

где

– среднее значение выборки

- квантиль нормального распределения с уровнем доверия

– стандартное отклонение выборки

– размер выборки

Значение зависит от выбранного уровня доверия и может быть найдено в таблице квантилей нормального распределения. Например, для уровня доверия 95% соответствующее значение равно 1,96. Ширина доверительного интервала зависит от размера выборки и величины стандартного отклонения. Чем больше размер выборки, тем меньше ширина интервала и тем выше точность оценки параметра. Чем больше стандартное отклонение, тем шире интервал и тем ниже точность оценки.

В нашей работе мы использовали метод бутстрэп (англ. bootstrap) для построения доверительных интервалов. Этот метод заключается в том, что мы создаем множество выборок с заменой из исходной выборки и вычисляем оптимальные гиперпараметры для каждой из них. Затем мы строим доверительный интервал на основе распределения оптимальных гиперпараметров. Этот метод позволяет улучшить точность оценки доверительного интервала и учитывать неопределенность в выборке.

Рассмотрим бутстрэповский алгоритм нахождения доверительного интервала:

Пусть дана выборка из генеральной совокупности, и требуется оценить параметр . Необходимо выбрать количество псевдовыборок, которые будут формироваться из элементов исходной выборки с возвращением (при формировании псевдовыборки элементы исходной выборки могут повторяться). Для каждой из псевдовыборок вычисляется псевдостатистика .

Псевдостатистики сортируются от меньшей к большей. Квантилями принимаются значения . С их помощью строится доверительный интервал.

* 1. **Исходные данные с месторождений**

Для необходимых исследований были предоставлены данные, полученные с помощью телеметрии на месторождениях (рисунок 2.1). В рамках задачи будет рассматриваться месторождение N. Диапазон данных включает замеры за несколько месяцев. Среди данных встречается множество пропущенных параметров и данных, которые не будут участвовать в исследовании (рисунки 2.2 и 2.3). Исходя из этого, перед началом проведения экспериментов необходимо провести предобработку и подготовку данных к работе.

Изображение выглядит как текст, снимок экрана, число, Параллельный

Автоматически созданное описание

Рисунок 2.1. Вид исходной таблицы с данными

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 2.2. Пропущенные данные по столбцам

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 2.3. Пропущенные данные по строкам

По результатам проведения предобработки, парсинга и чистки на выходе получаются следующие данные (рисунок 2.4).

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 2.4. Данные после предобработки и парсинга в Pandas DataFrame

Здесь, как можно увидеть, были удалены не используемые в исследовании столбцы данных, произведен поиск и удаление строк с пропущенными значениями параметров и произведен парсинг из формата Excel в Pandas DataFrame [14].

**ГЛАВА 3. ПРАКТИЧЕСКАЯ ЧАСТЬ.**

**3.1. Алгоритм программы для нахождения оптимальных гиперпараметров**

Изучая данные, представленные в теоретической части, была написана программа, которая анализирует входные данные, строит модель машинного обучения добывающей скважины и находит доверительные интервалы гиперпараметров с помощью различных методов оптимизации. Алгоритм программы имеет следующий характер:

1. Импортирование необходимых библиотек и загрузка данных из файлов Excel.
2. Проведение парсинга и предобработки исходных данных.
3. Определение целевой переменной (забойное давление) и признаков (обводненность, газовый фактор, устьевое давление и дебит).
4. Разделение данных на тренировочную и тестовую выборки, выбор метрики для оценивания полученных результатов и задание интервалов гиперпараметров.
5. Применение методов Grid Search, Random Search, CMA-ES от Optuna и TPE от Hyperopt для нахождения лучших параметров модели [15].
6. Нахождение доверительных интервалов гиперпараметров для каждого из методов.
7. Вывод графиков сравнения точности, затраченного времени и доверительных интервалов для каждого из методов.

Основные функции программы:

Grid\_search: использует метод полного перебора для нахождения лучших гиперпараметров модели Gradient Boosting Regressor. Она создает сетку гиперпараметров, задаваемую вручную, и оценивает модель для каждой комбинации гиперпараметров с помощью кросс-валидации.

Random\_search: применяет случайный поиск для отыскания наилучших гиперпараметров модели.

CMA-ES\_search: применяет метод CMA-ES из библиотеки Optuna для автоматической настройки гиперпараметров модели машинного обучения. Определяет пространство поиска гиперпараметров и использует алгоритмы оптимизации библиотеки Optuna для нахождения лучших гиперпараметров. Значения, найденные с помощью этого метода, используются для обучения модели на обучающих данных и генерации предсказаний на тестовых данных.

TPE\_search: задает диапазон возможных значений гиперпараметров и использует алгоритм TPE оптимизации из библиотеки Hyperopt для поиска наилучших гиперпараметров. После нахождения оптимальных гиперпараметров модель обучается на обучающих данных и используется для предсказания на тестовых данных.

Plot\_prediction\_results: визуализирует результаты предсказаний модели, обученной с помощью различных методов настройки гиперпараметров.

Bootstrap\_confidence\_intervals: использует метод бутстреппинга для оценки качества гиперпараметров, найденных различными методами настройки. Она генерирует множество обучающих выборок с заменой и оценивает качество гиперпараметров, найденных каждым методом, на каждой выборке. Затем она строит доверительные интервалы для каждого гиперпараметра и визуализирует их с помощью графиков.

**3.2. Результаты работы программы**

В результате работы программы были найдены оптимальные значения гиперпараметров модели и произведен поиск доверительных интервалов гиперпараметров для каждого из методов. Также были получены гистограммы, визуализирующие сравнение точности предсказаний каждого из методов и графики, показывающие сопоставление истинных значений целевой переменной и предсказанных.

На рисунках 3.1–3.10 представлены графики зависимости предсказанного значения целевой переменной от истинного значения для пяти скважин. Программа показывает результаты предсказаний для каждого из четырех методов.

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Рисунок 3.1. График предсказанного забойного давления от истинного для скважины 1

Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма

Автоматически созданное описание

Рисунок 3.2. Гистограмма распределения предсказанных и истинных значений скважины 1

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Автоматически созданное описание

Рисунок 3.3. График предсказанного забойного давления от истинного для скважины 2

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 3.4. Гистограмма распределения предсказанных и истинных значений скважины 2

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Автоматически созданное описание

Рисунок 3.5. График предсказанного забойного давления от истинного для скважины 3

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Рисунок 3.6. Гистограмма распределения предсказанных и истинных значений скважины 3

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Автоматически созданное описание

Рисунок 3.7. График предсказанного забойного давления от истинного для скважины 4

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 3.8. Гистограмма распределения предсказанных и истинных значений скважины 4

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Рисунок 3.9. График предсказанного забойного давления от истинного для скважины 5

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 3.10. Гистограмма распределения предсказанных и истинных значений скважины 5

Исходя из полученных графиков и гистограмм, можно сделать вывод, что все используемые нами методы позволяют получить достаточно точные значения целевой переменной, а именно забойного давления.

Теперь перейдем к таблице 3.1, в которой представлено сравнение времени, затраченное каждым из методов на нахождение оптимальных гиперпараметров для каждой скважины [2], и точности, с которой данные методы предсказывают значение забойного давления на тестовой выборке. Точность в данном случае вычисляется с помощью кастомной функции, которая принимает на вход два массива: фактические значения целевой переменной (забойное давление) и предсказанные значения, полученные с помощью одного из методов оптимизации. Затем, в зависимости от величины фактического значения целевой переменной, функция проверяет, удовлетворяет ли разность между фактическим и предсказанным значениями определенному критерию.

Таблица 3.1. Сравнение точности и времени работы для каждого из методов

|  |  |  |  |
| --- | --- | --- | --- |
| **№ скважины** | **Метод оптимизации** | **Затраченное время, сек** | **Точность** |
| 1 | GridSearch | 31,61348486 | 0,893617 |
| RandomSearch | 3,256240368 | 0,893617 |
| CMA-ES | 11,73958 | 0,893617 |
| TPE | 14,08993 | 0,893617 |
| 2 | GridSearch | 21,35063 | 0,8 |
| RandomSearch | 2,768168 | 0,76 |
| CMA-ES | 9,731938 | 0,8 |
| TPE | 13,25955 | 0,8 |
| 3 | GridSearch | 20,40328 | 0,84 |
| RandomSearch | 2,249022 | 0,74 |
| CMA-ES | 10,47003 | 0,79 |
| TPE | 10,34782 | 0,74 |
| 4 | GridSearch | 19,54265 | 0,952381 |
| RandomSearch | 2,08997 | 0,904762 |
| CMA-ES | 9,858635 | 0,952381 |
| TPE | 11,87252 | 0,952381 |
| 5 | GridSearch | 20,46619 | 0,909091 |
| RandomSearch | 2,210112 | 0,863636 |
| CMA-ES | 10,95232 | 0,909091 |
| TPE | 10,51911 | 0,863636 |

Исходя из полученных данных видно, что практически все методы оптимизации показали достаточно хорошие результаты в терминах точности. Как и требовалось ожидать, GridSearch показал наивысшую точность во всех экспериментах благодаря тому, что он использует в своей механике полный перебор всех возможным комбинаций гиперпараметров и находит лучший из них, но он является и самым затратным по времени. Наименее точным оказался RandomSearch, в четырех случаях из пяти он показал худшее значение точности, но, с другой стороны, данный метод занял первое место среди всех по затраченному времени. Методы CMA-ES и TPE показали схожие результаты в терминах точности и времени, занятого для оптимизации.

Далее рассмотрим таблицу 3.1. Здесь представлены найденные доверительные интервалы гиперпараметров для каждого из методов.

Таблица 3.1. Полученные доверительные интервалы

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ скважины** | **Метод оптимизации** | **learning\_rate** | **max\_depth** | **min\_samples\_split** | **n\_estimators** |
| 1 | GridSearch | [0.05, 0.1] | [4, 6] | [4, 5] | [10, 75] |
| RandomSearch | [0.1, 0.2] | [4, 10] | [4, 8] | [50, 125] |
| CMA-ES | [0.05, 0.2] | [4, 8] | [4, 7] | [10, 100] |
| TPE | [0.1, 0.2] | [4, 7] | [4, 7] | [10, 50] |
| 2 | GridSearch | [0.05, 0.1] | [4, 6] | [4, 8] | [40, 100] |
| RandomSearch | [0.1, 0.2] | [4, 10] | [5, 8] | [50, 125] |
| CMA-ES | [0.05, 0.1] | [5, 8] | [4, 8] | [40, 100] |
| TPE | [0.1, 0.2] | [4, 5] | [4, 4] | [10, 20] |
| 3 | GridSearch | [0.05, 0.1] | [4, 6] | [4, 7] | [10, 50] |
| RandomSearch | [0.05, 0.2] | [4, 10] | [5, 8] | [30, 100] |
| CMA-ES | [0.05, 0.1] | [7, 9] | [5, 8] | [20, 100] |
| TPE | [0.05, 0.2] | [4, 7] | [4, 7] | [10, 50] |
| 4 | GridSearch | [0.05, 0.2] | [5, 6] | [4, 6] | [20, 75] |
| RandomSearch | [0.05, 0.2] | [5, 9] | [4, 6] | [20, 75] |
| CMA-ES | [0.05, 0.2] | [5, 10] | [4, 8] | [10, 75] |
| TPE | [0.05, 0.2] | [4, 7] | [4, 6] | [10, 40] |
| 5 | GridSearch | [0.05, 0.2] | [4, 6] | [4, 6] | [10, 100] |
| RandomSearch | [0.05, 0.2] | [5, 10] | [4, 7] | [20, 125] |
| CMA-ES | [0.1, 0.2] | [4, 9] | [4, 8] | [10, 50] |
| TPE | [0.05, 0.2] | [4, 7] | [4, 6] | [10, 50] |

После анализа полученных доверительных интервалов для данных скважин можно сделать несколько выводов:

1. GridSearch является наиболее подходящим методом оптимизации в рамках данной задачи, так как он позволяет получить достаточно качественные модели за разумное количество итераций.
2. RandomSearch не является эффективным в рамках данной задачи, из-за того, что он не улучшает качество модели по сравнению с GridSearch.
3. CMA-ES и TPE являются перспективными методами оптимизации для данной задачи, так как они позволяют получить более качественные модели, чем GridSearch и RandomSearch, за меньшее количество итераций. Однако, для их использования требуются более мощные вычислительные ресурсы.

Подводя итог, рассматривая полученные интервалы моделей, оптимизированных с помощью GridSearch, RandomSearch, CMA-ES и TPE, можно сформировать доверительные интервалы для месторождения N следующим образом:

* learning\_rate: [0.05, 0.2]
* max\_depth: [4, 9]
* min\_samples\_split: [4, 8]
* n\_estimators: [10, 100]

При этом, на формирование доверительных интервалов для месторождения GridSearch будет влиять больше, чем другие методы оптимизации, а RandomSearch практически не будет влиять.

**ЗАКЛЮЧЕНИЕ**

В рамках проведенной работы была изучена задача нахождения оптимальных гиперпараметров для моделей машинного обучения, применяемых в нефтедобывающей отрасли. Были рассмотрены различные методы оптимизации гиперпараметров, такие как GridSearch, RandomSearch, CMA-ES и TPE.

Результаты, полученные путем сравнения этих методов на нескольких скважинах, показали, что все они способны найти гиперпараметры, обеспечивающие хорошую точность модели. Однако, время, затрачиваемое на оптимизацию, значительно различается. Кроме того, важно обратить внимание на доверительные интервалы, полученные в результате оптимизации. Их анализ позволяет оценить стабильность и надежность модели, а также выявить гиперпараметры, наиболее существенно влияющие на ее качество.

Важным моментом, который стоит отметить, является то, что найденные доверительные интервалы гиперпараметров могут быть использованы не только в рамках одной задачи, но и для оптимизации других моделей машинного обучения, применяемых в нефтедобывающей отрасли. Таким образом, проведенная работа позволяет не только улучшить качество и сократить время на разработку конкретной модели, но и иметь общие рекомендации по оптимизации гиперпараметров для различных моделей, используемых в отрасли. Кроме того, анализ доверительных интервалов позволяет выявить наиболее существенные гиперпараметры, на которые следует обращать особое внимание при оптимизации, что также способствует улучшению качества моделей.

**СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ**

1. Крец В. Г., Шадрина А. В. Основы нефтегазового дела. Учебное пособие. – Томск: Изд-во Томского политехнического университета, 2011. – 138 с.
2. Любанович Билл. Простой Python. Современный стиль программирования. 2-е изд. — СПб.: Питер, 2021 — 275 с.:
3. Минханов, И. Ф., Разработка нефтяных и газовых месторождений: учеб. пособие для вузов / И. Ф. Минханов, С. А. Долгих, М. А. Варфоломеев; Казанский федеральный университет. – Казань, 2019. – 96с.
4. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — СПб.: «Питер», 2018. — с. 331-375.
5. Рашка С., Мирджалили В. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow. – СПб.: ООО «Диалектика», 2019.
6. Фролов А. Н. Краткий курс теории вероятностей и математической статистики: учебное пособие для СПО. — СПб.: Лань, 2021. — С. 189. — 316 с.
7. Хеллман Даг. Стандартная библиотека Python 3: справочник с примерами, 2-е изд. СПб.: ООО “Диалектика”, 2019. — 1376 с.
8. Захаров Л.А., Мартюшев Д.А., Пономарева И.Н. Прогнозирование динамического пластового давления методами искусственного интеллекта // Записки Горного института. 2022.
9. Firouzi M., Rathnayake S. Prediction of the flowing bottom-hole pressure using advanced data analytics //Asia Pacific Unconventional Resources Technology Conference, Brisbane, Australia, 18-19 November 2019.–С. 735-742.
10. 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.
11. Ignatov D., Sinkov K., Spesivtsev P., Vrabie I. Tree-Based Ensembles for Predicting the Bottomhole Pressure of Oil and Gas Well Flows: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers.
12. Mohammadpoor M., Shahbazi Kh., Torabi F. and Qazvini A. A New Methodology for Prediction of Bottomhole Flowing Pressure in Vertical Multiphase Flow in Iranian Oil Fields Using Artificial Neural Networks (ANNs)//SPE Latin American & Caribbean Petroleum Engineering Conference. – OnePetro, 2010.
13. 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.
14. Документация библиотеки Pandas. - https://pandas.pydata.org/docs/
15. Документация библиотеки scikit-learn. - <https://scikit-learn.org/stable/>