НАХОЖДЕНИЕ ОПТИМАЛЬНЫХ ГИПЕРПАРАМЕТРОВ ДЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДОБЫВАЮЩИХ СКВАЖИН

Выпускная квалификационная работа

Студент: Кукуев А.И.

Руководитель: Симонов М.В., старший преподаватель ВШТМиМФ

Консультант: Печко К.А., главный специалист НОЦ «Газпромнефть-Политех»

Актуальность

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. 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.

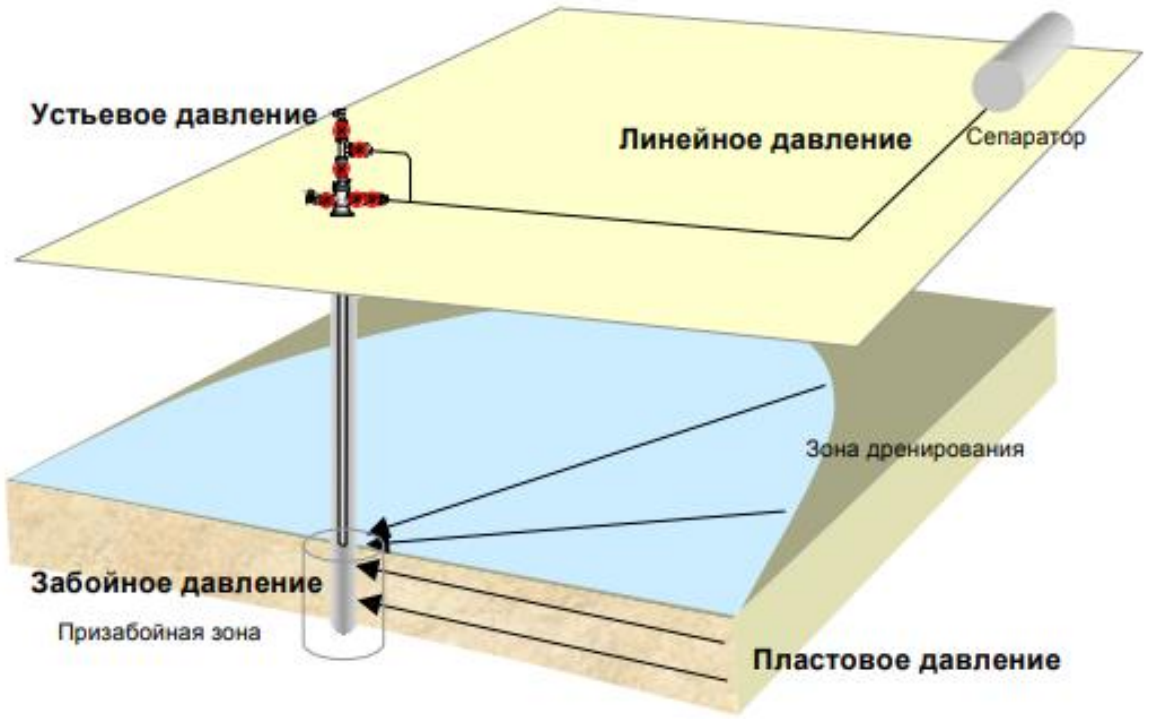
Цель

Нахождение доверительных интервалов гиперпараметров для моделей машинного обучения добывающих скважин.

Задачи:

- Создать модель машинного обучения для добывающих скважин на языке Python, предсказывающей значение целевой переменной относительно параметров, регистрируемых телеметрией на месторождениях
- Проверить эффективность созданной модели на основе данных, полученных с месторождений
- Провести сбор и анализ гиперпараметров, используемых в модели, и сформировать выборки их значений
- Сделать статический анализ и построить доверительные интервалы для гиперпараметров модели

Обзор используемых в исследовании параметров добывающей скважины



Целевая переменная – **забойное давление**

Параметры:

- Дебит скважины
- Устьевое давление
- Газовый фактор
- Обводненность скважины

Сведения о параметрах скважины

- $\boldsymbol{P_з = \rho_ж g H + P_у}$ – формула для расчета забойного давления (без учета трения)

$P_з$ – давление на забое скважины, Па;

H – высота столба жидкости в скважине, м;

$\rho_ж$ – плотность жидкости, кг/м3;

$P_у$ – давление на устье скважины, Па.

Сведения о параметрах скважины

- $\boldsymbol{\frac{\partial P}{dx} = -\frac{f\rho v^2}{2D}}$ – формула Дарси-Вейсбаха (с учетом потерь на трение)

$\frac{\partial P}{dx}$ – градиент давления, Па/м;

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

ρ – плотность флюида, кг/м3;

v – скорость флюида, м/с;

D – диаметр скважины, м.

Сведения о параметрах скважины

- $\boldsymbol{Q = \frac{HV}{H_д - H_{ст}}}$ – формула расчета дебита нефтяной скважины

Q – дебит;

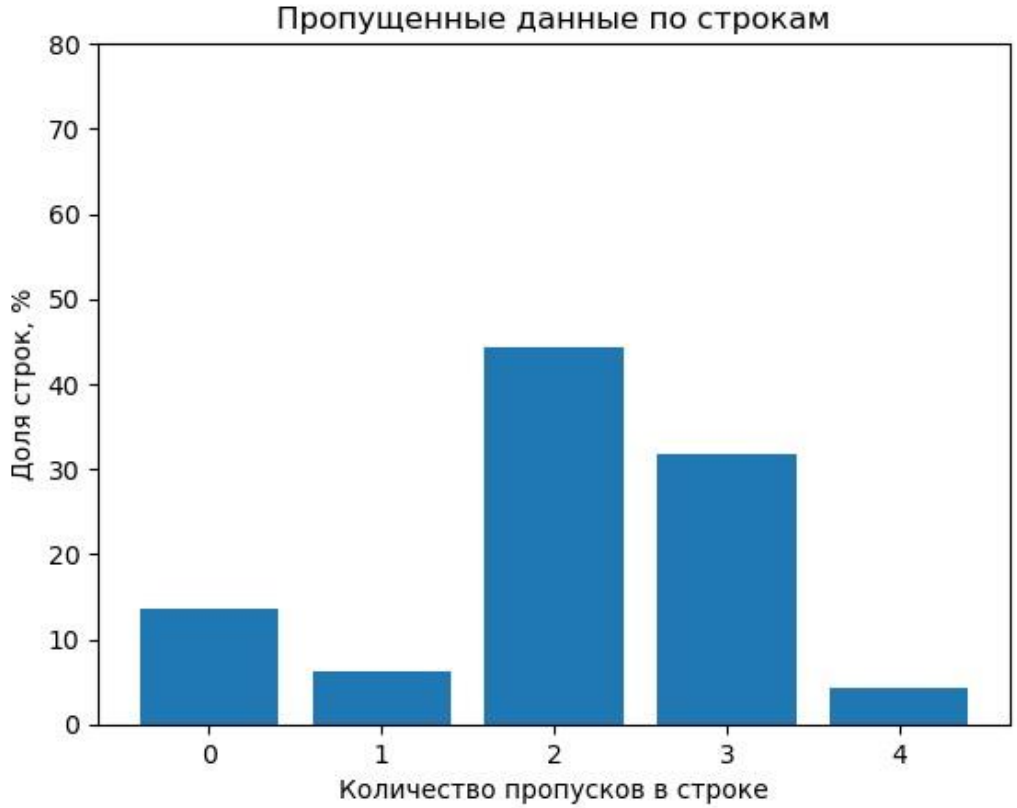
H – высота столба жидкости в скважине, м;

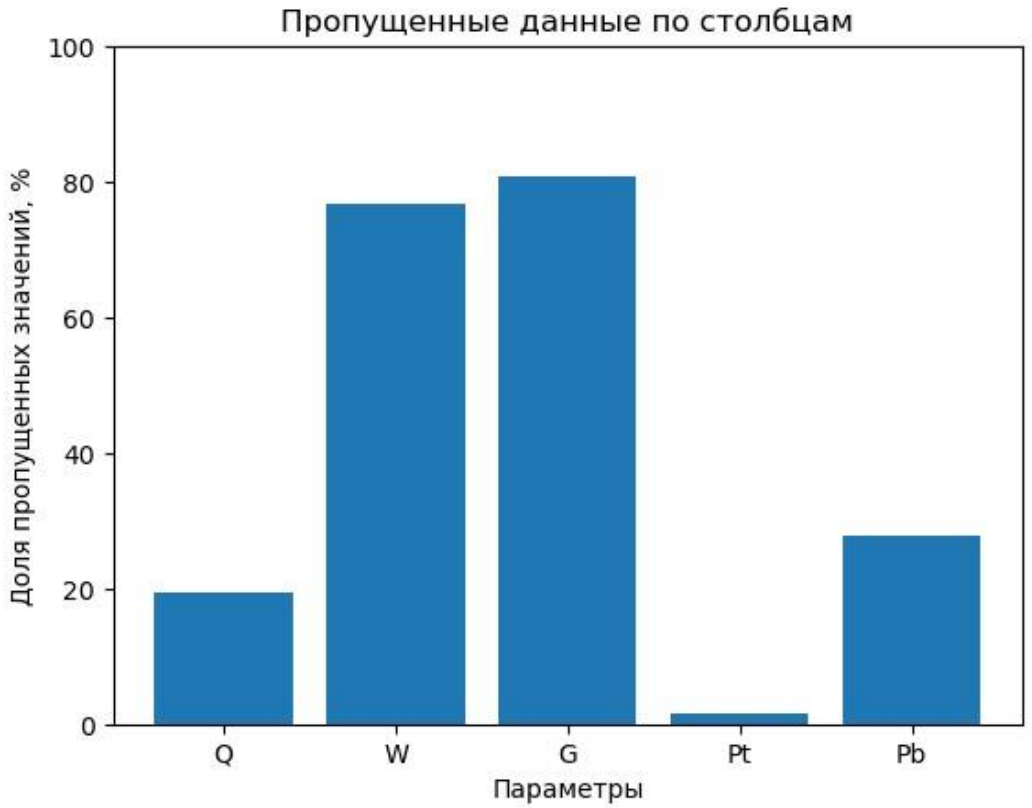
V – производительность насоса;

$H_{ст}$ – статический уровень, расстояние от начала подземных вод до первых слоёв почвы;

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

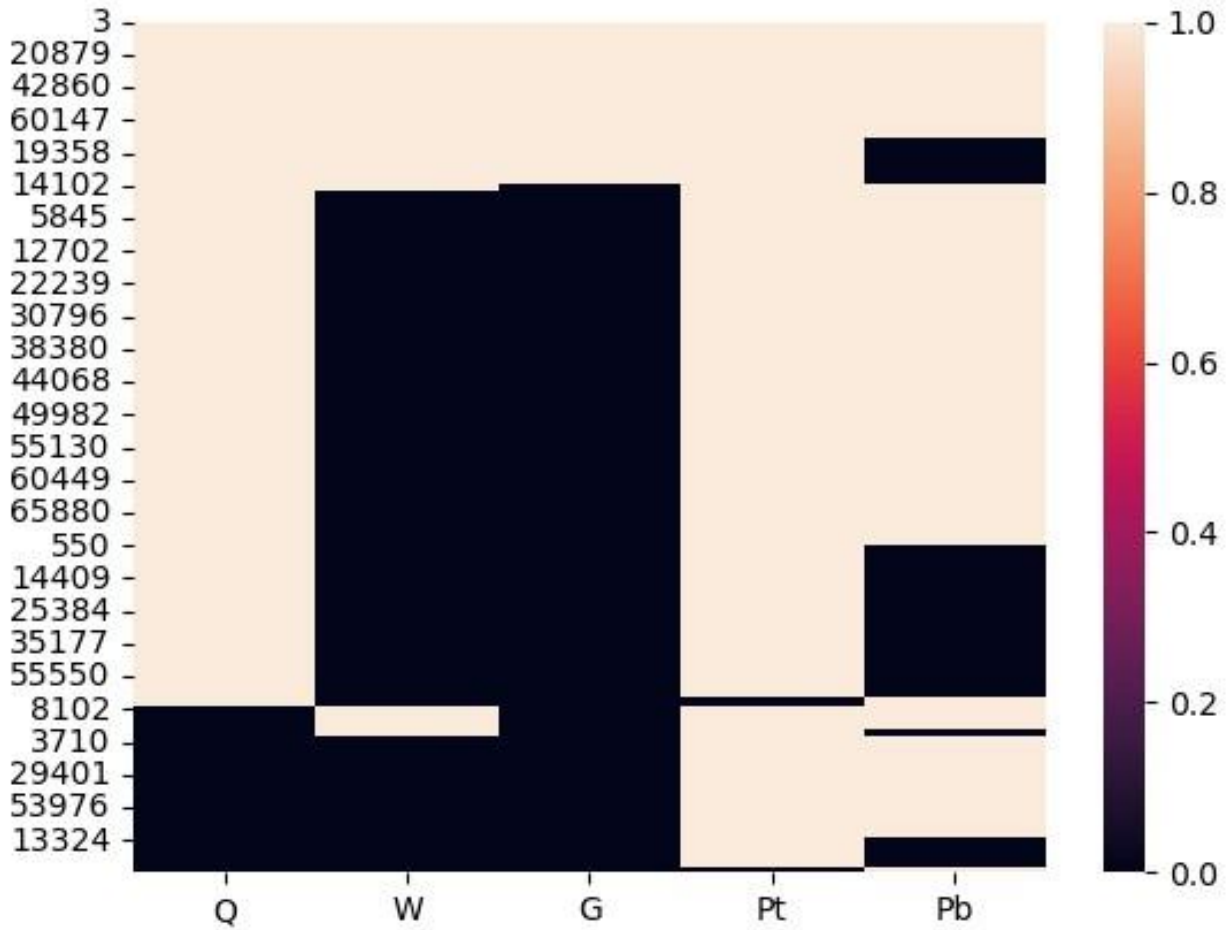
Обзор исходных данных с месторождения

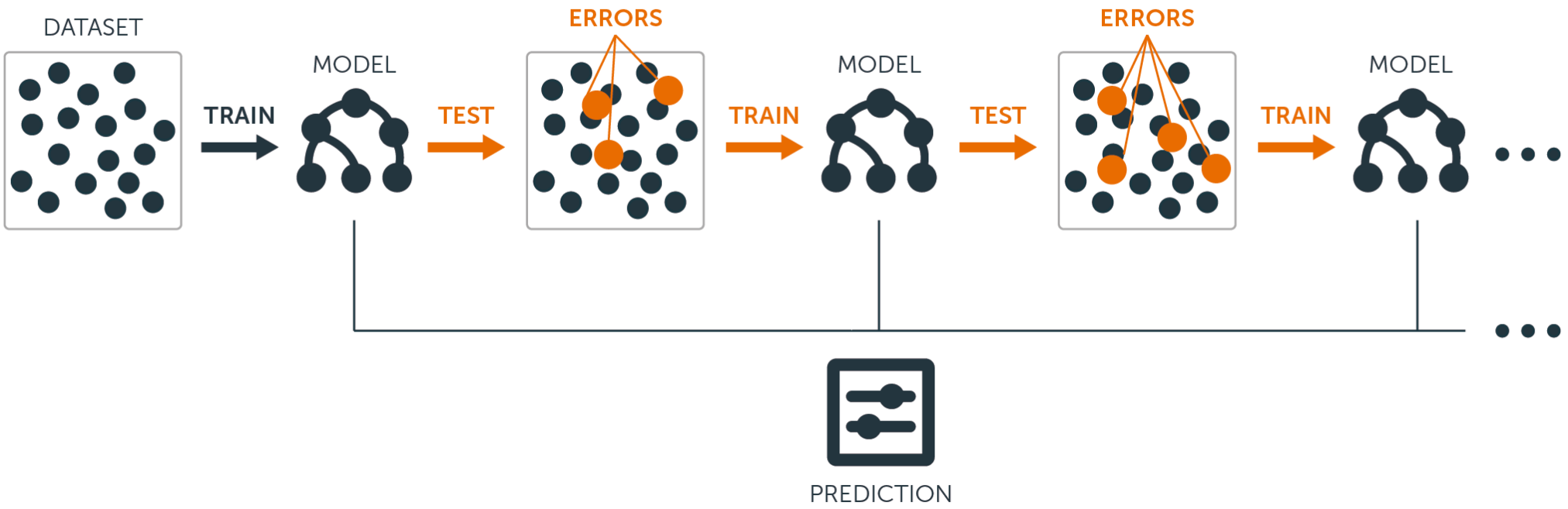




Обзор исходных данных с месторождения

Из распределения видно, что по большому количеству строк отсутствуют одновременно данные по ГФ и обводненности





Модель машинного обучения

В данной работе используется модель градиентного бустинга от sklearn. Она является моделью ансамблевого обучения, которая объединяет множество слабых моделей для создания одной сильной.

Методы оптимизации

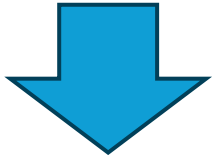
Метод оптимизации	Grid Search	Random search	CmaEsSampler (CMA-ES) из библиотеки Optuna	Tree-structured Parzen Estimator (TPE) из библиотеки Hyperopt
Преимущества	- Простота реализации - Высокая точность результатов	- Меньшая вычислительная сложность по сравнению с Grid Search - Легко поддается параллельной обработке	- Наличие встроенных визуализаций - Имеет поддержку распределенных вычислений	- Использует алгоритм деревьев решений - Возможность сохранять и загружать результаты оптимизации -
Недостатки	- Высокая вычислительная сложность - Неэффективен при большом количестве гиперпараметров	- Отсутствие гарантий нахождения - Необходимость в большом количестве испытаний	- Высокие требования к ресурсам - Сложный синтаксис и недостаток документации	- Более сложен в реализации - Чувствителен к начальным условиям

Алгоритм Grid Search

Определение пространства гиперпараметров

Создание сетки

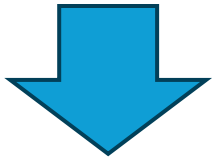
Оценка качества модели



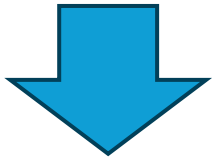
Выбор оптимальных гиперпараметров

Алгоритм Random Search

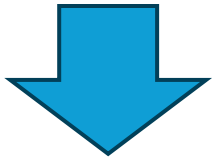
Определение пространства гиперпараметров



Случайный выбор значений



Оценка качества модели



Повторение шагов 2 и 3

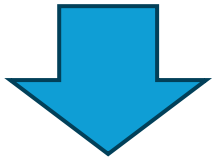
Выбор оптимальных гиперпараметров

Алгоритм CMA-ES

Инициализация и оценка качества

Отбор и обновление ковариационной матрицы

$$C_t = (1 - c_{cov})C_{t-1} + c_{cov}\frac{1}{p_s}\sum_{i=1}^{m}\left(\lambda_i - \bar{\lambda}\right)\left(\lambda_i - \bar{\lambda}\right)^T$$



Обновление вектора шага: $\sigma_t = \sigma_{t-1}\, exp\left(\frac{c_s}{d_s}\left(\frac{|p_s|}{E|p_s|} - 1\right)\right)$

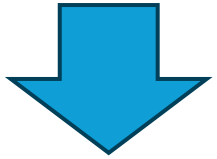
Обновление популяции и проверка условий остановки:

$$\lambda_i = \bar{\lambda} + \sigma_t N(0, C_t)$$

Алгоритм TPE

Инициализация и оценка качества

$f(\lambda)$, где λ – вектор гиперпараметров



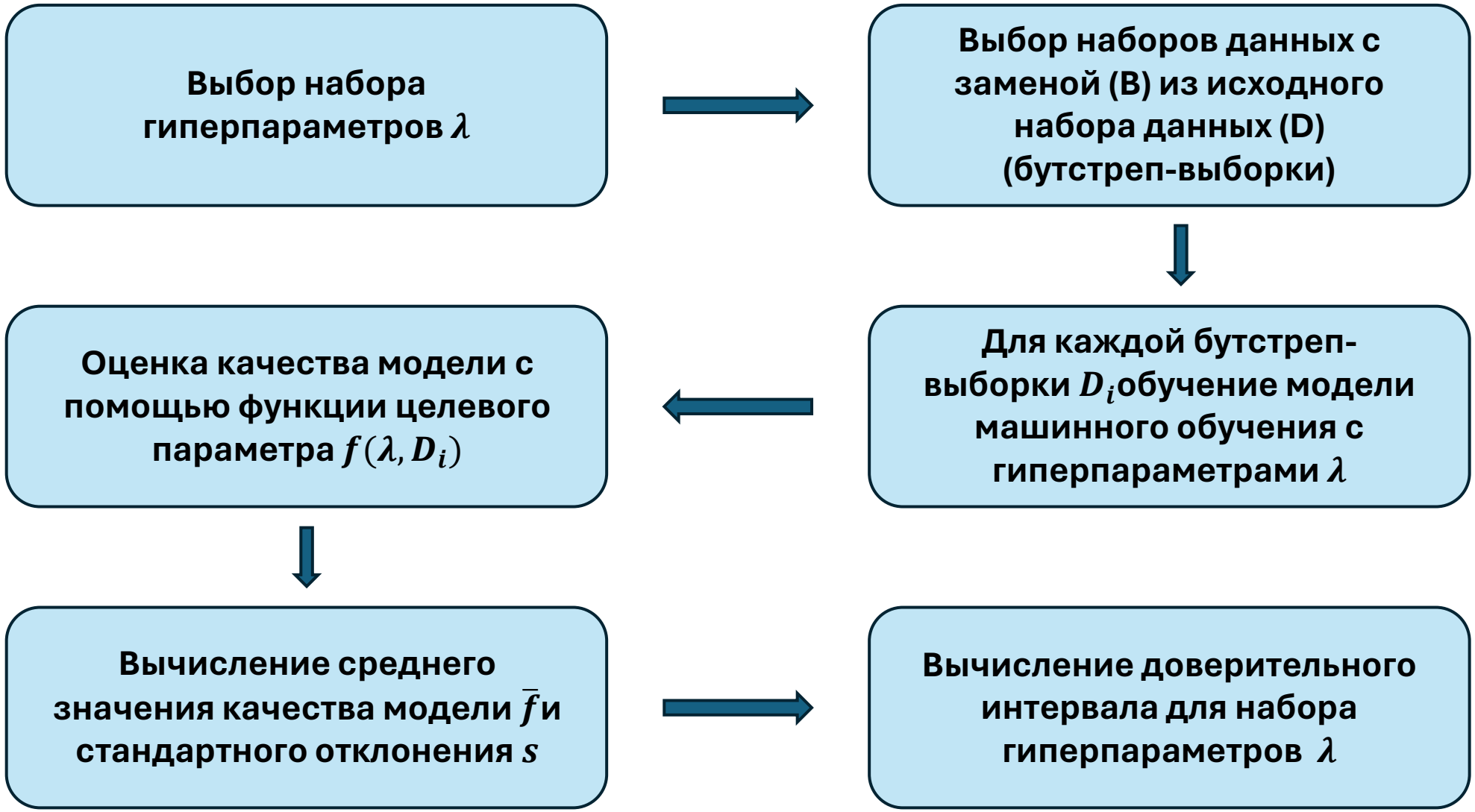
Построение дерева решений

$$l(\lambda) = \int_{-\infty}^{f(\lambda)} p(y)dy,\ g(\lambda) = \int_{f(\lambda)}^{\infty} p(y)dy,\ p(\lambda) = \frac{l(\lambda)}{l(\lambda)+g(\lambda)}$$

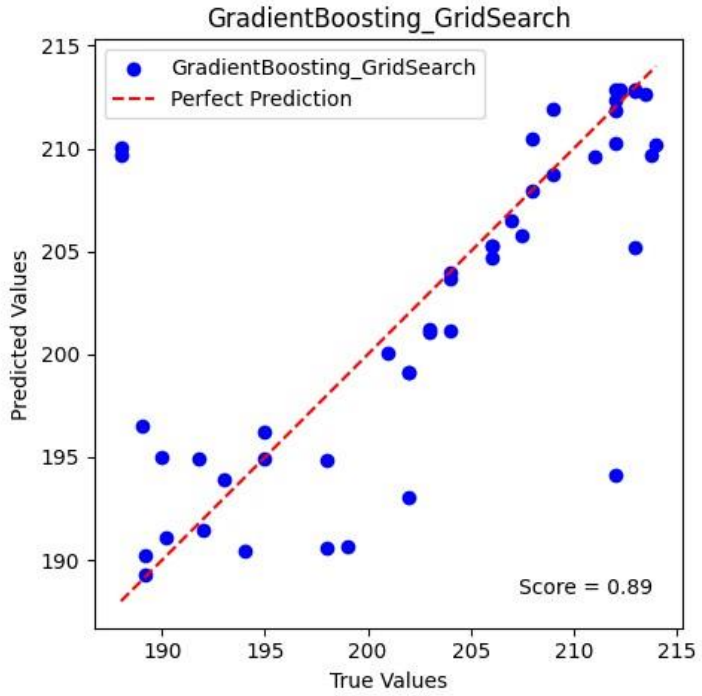
Выбор гиперпараметров: $\lambda_{t+1} = argmax\ \frac{p(\lambda)}{l(\lambda)+g(\lambda)}$

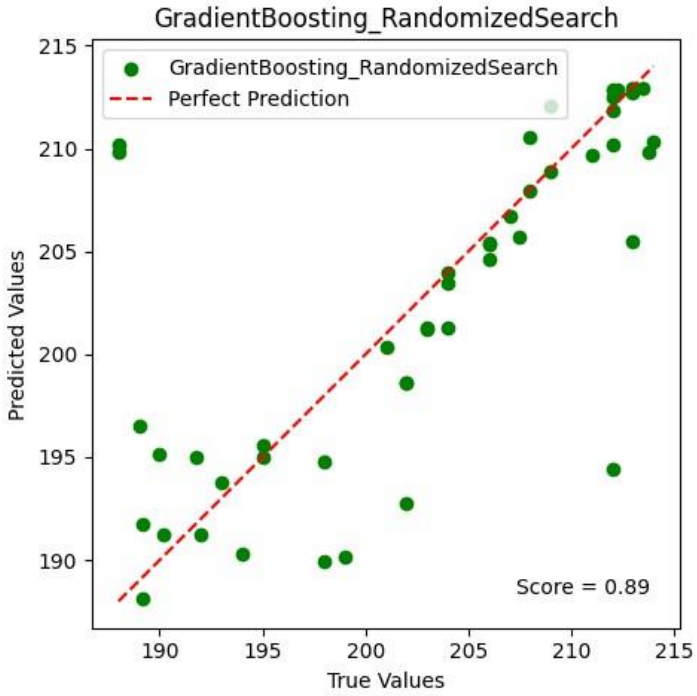
Проверка условий остановки

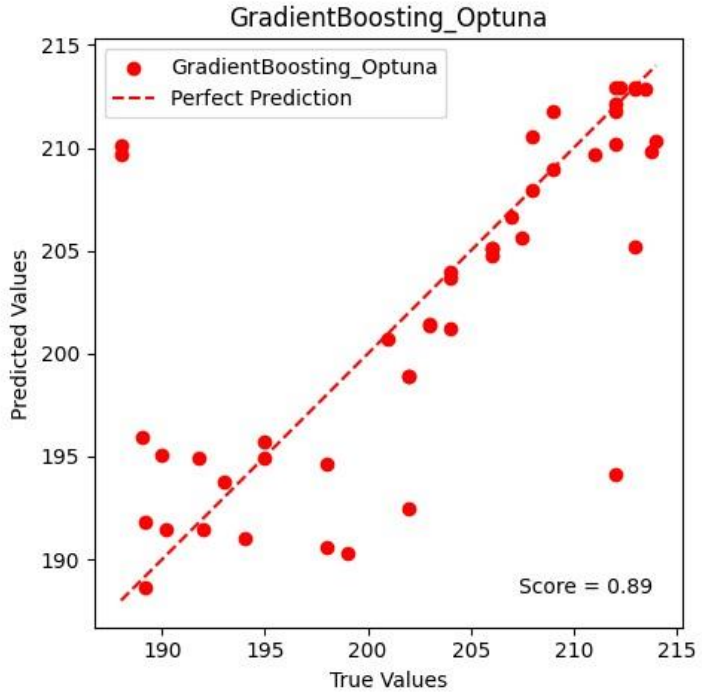
Метод bootstrap

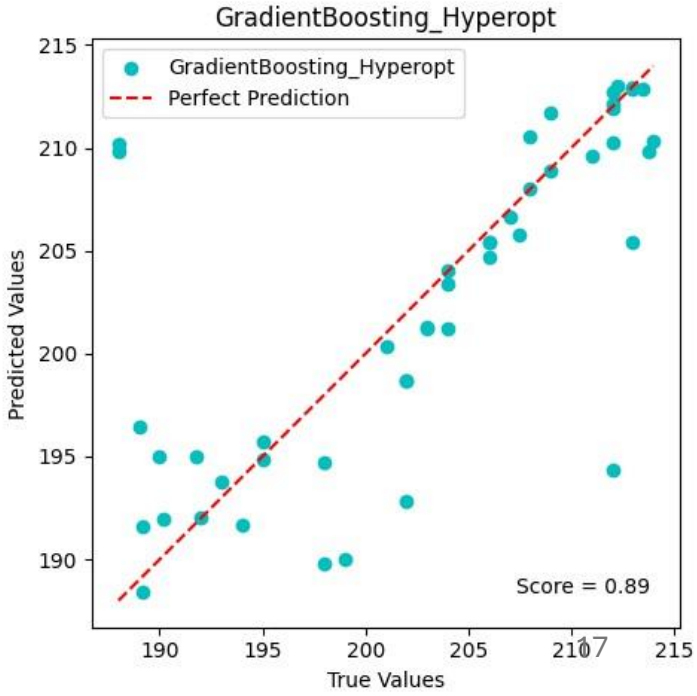


Кроссплот предсказанного забойного давления от истинного для скважины 1

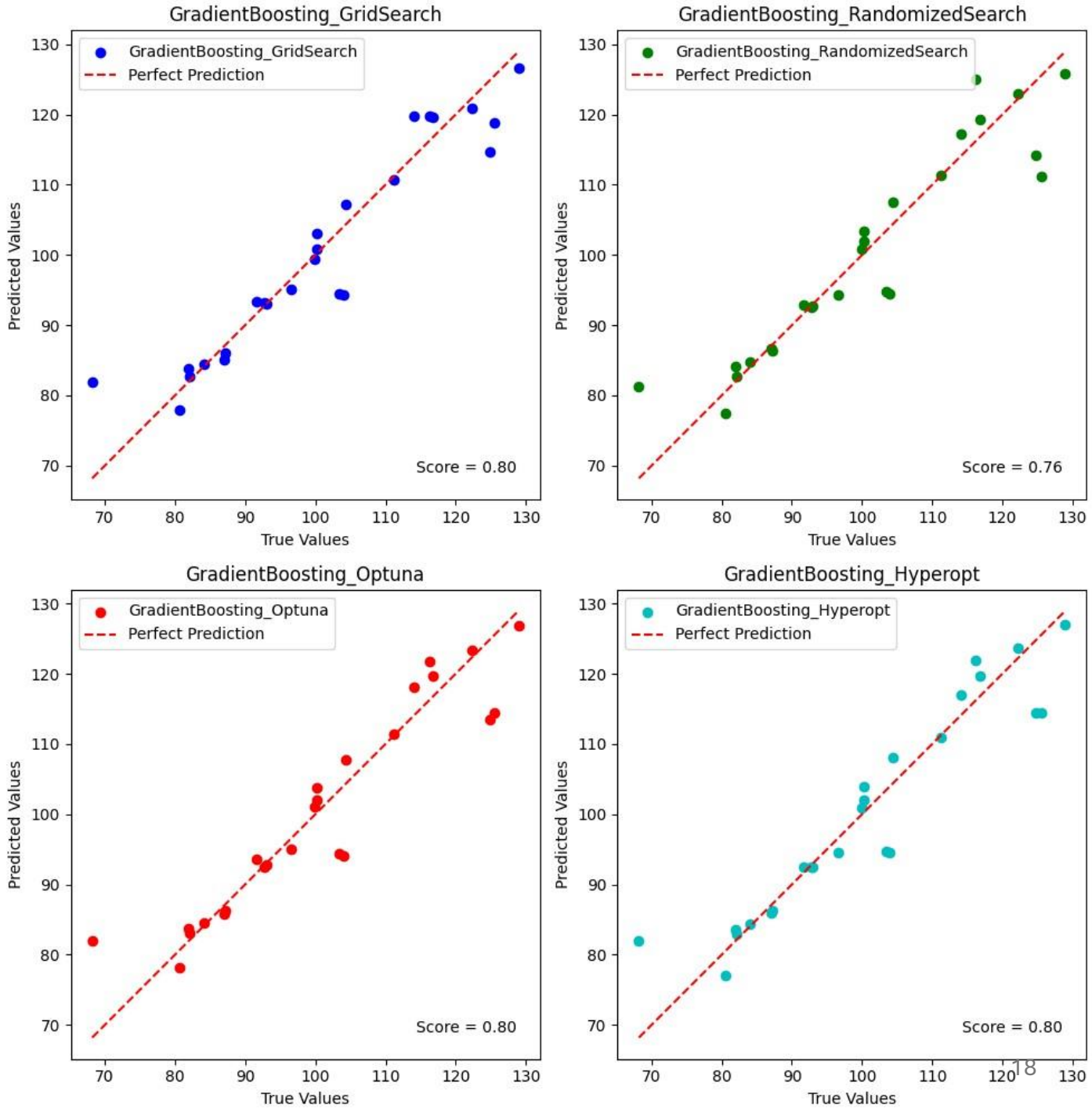




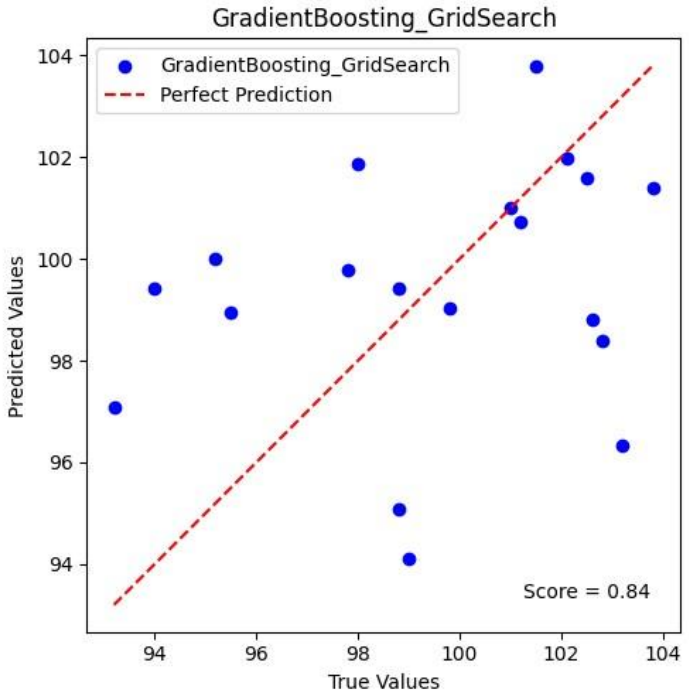


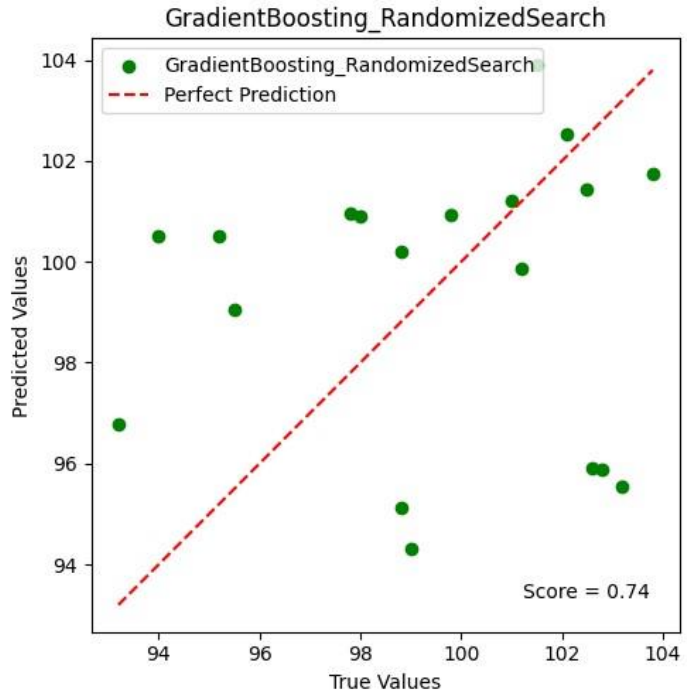


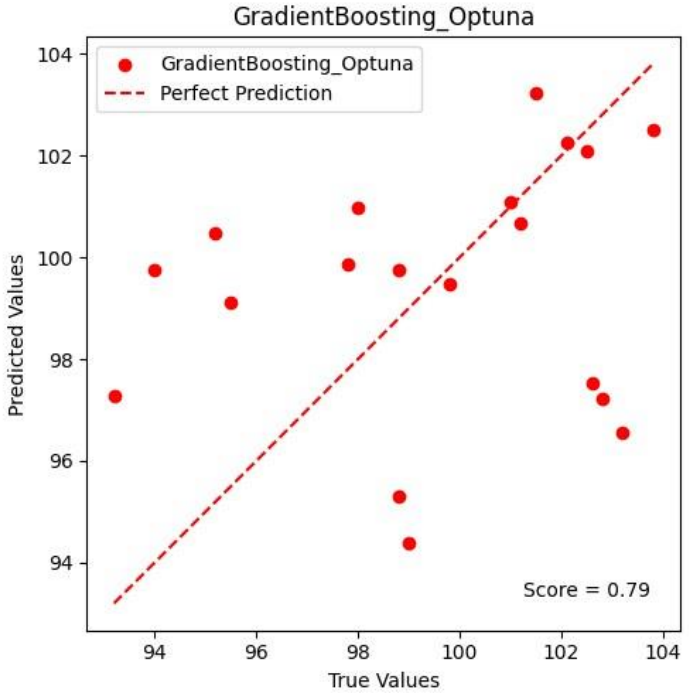
Кроссплот предсказанного забойного давления от истинного для скважины 2

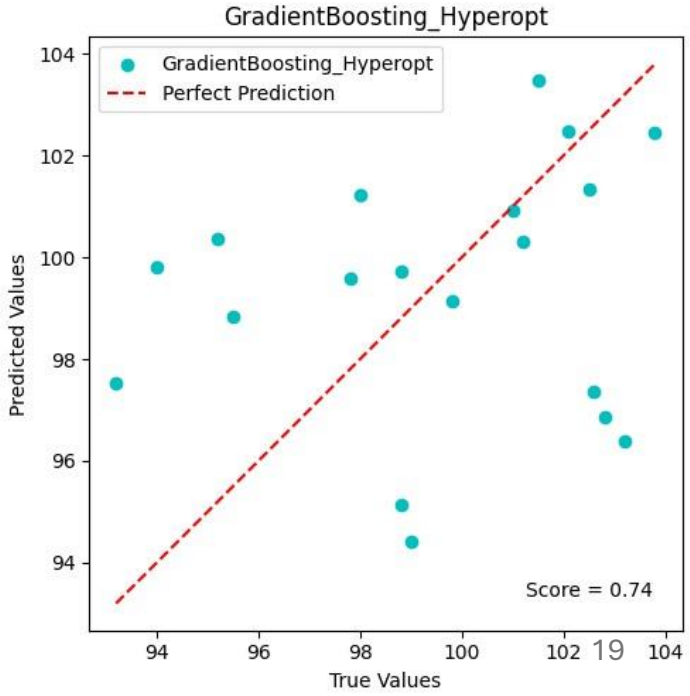


Кроссплот предсказанного забойного давления от истинного для скважины 3

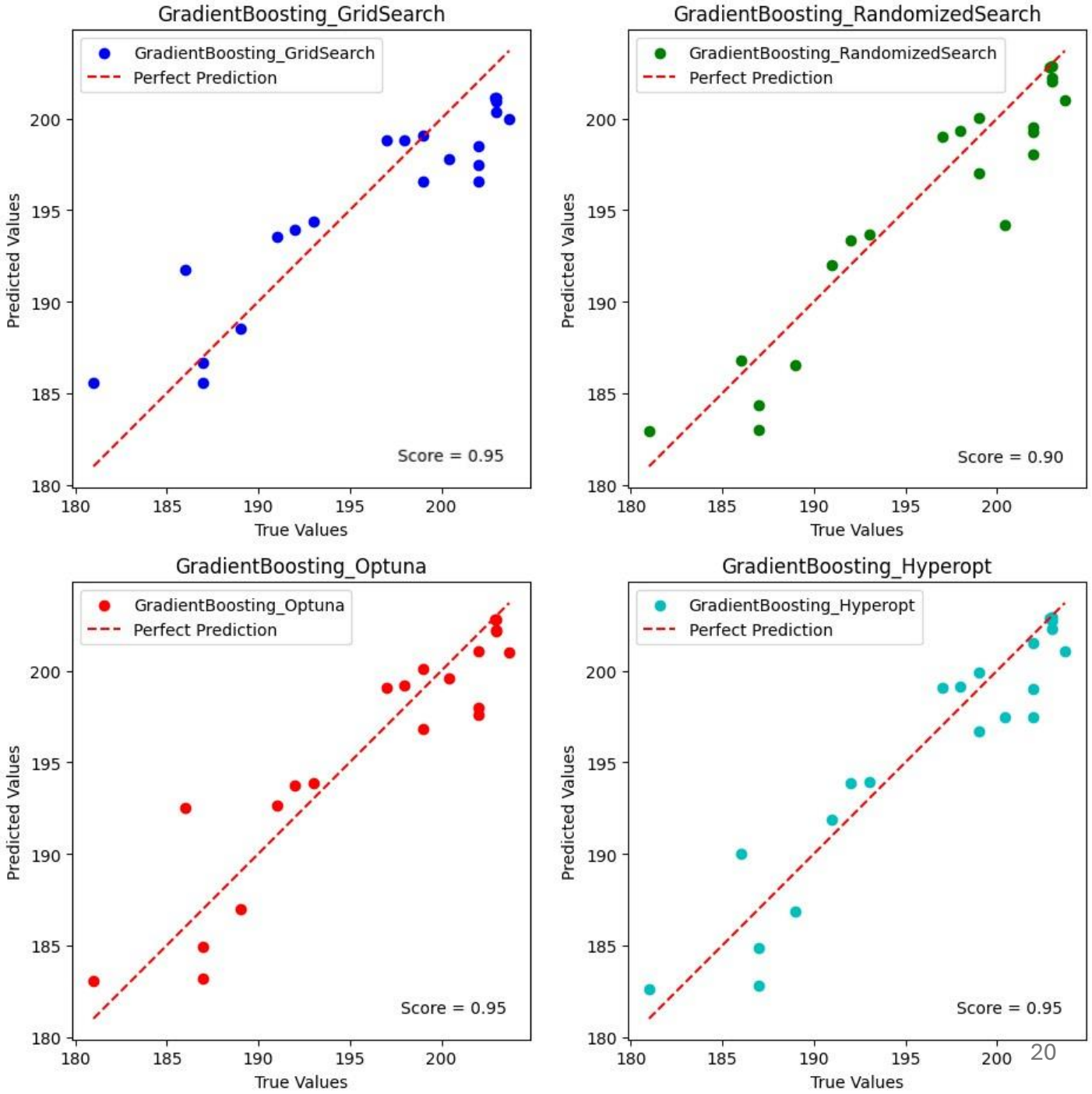




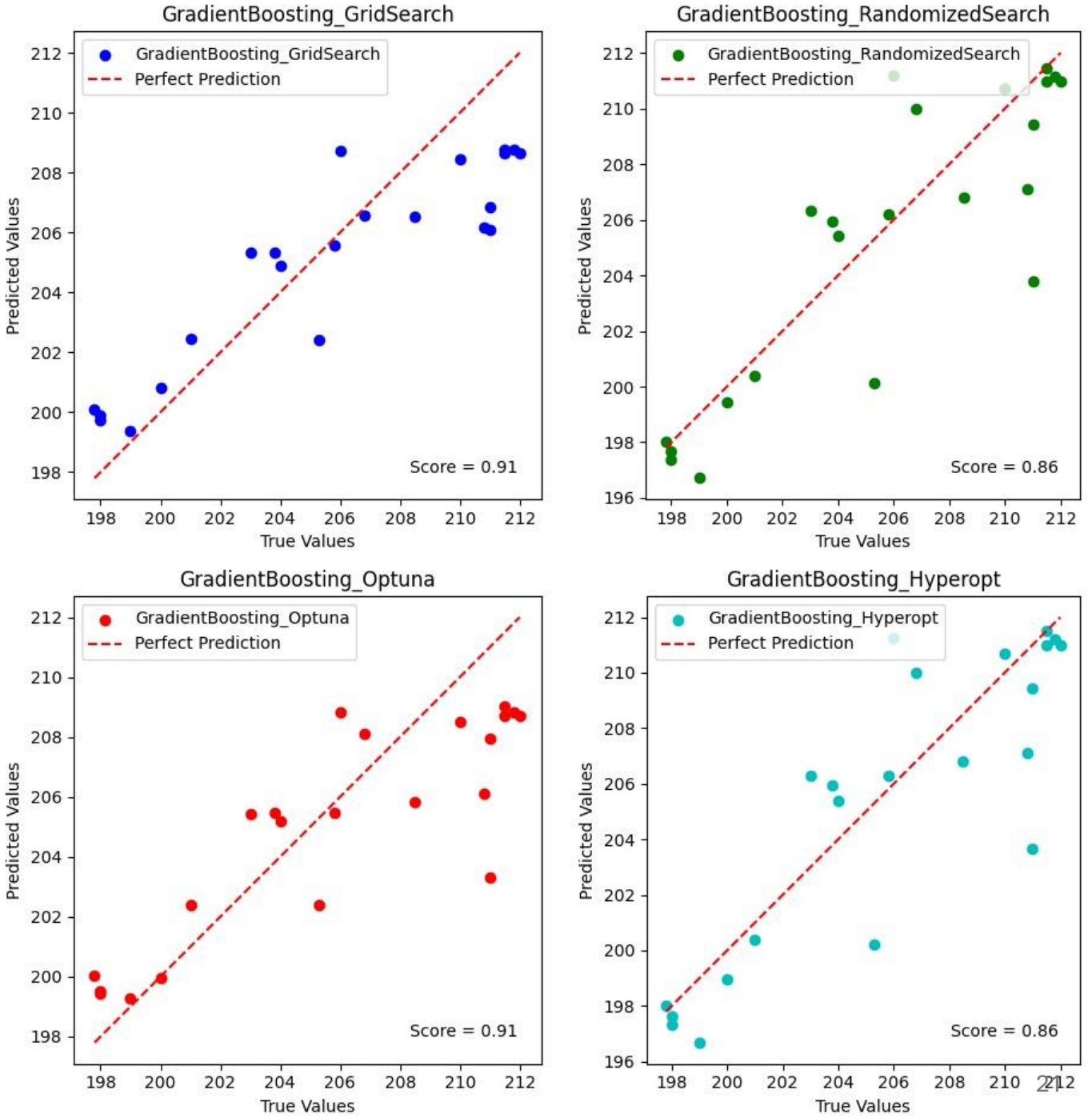




Кроссплот предсказанного забойного давления от истинного для скважины 4



Кроссплот предсказанного забойного давления от истинного для скважины 5



№ СКВАЖИНЫ	МЕТОД ОПТИМИЗАЦИИ	ЗАТРАЧЕННОЕ ВРЕМЯ, СЕК	ТОЧНОСТЬ
1	GridSearch	31,61	0,89
	RandomSearch	3,25	0,89
	CMA-ES	11,74	0,89
	TPE	14,1	0,89
2	GridSearch	21,35	0,80
	RandomSearch	2,77	0,76
	CMA-ES	9,73	0,80
	TPE	13,26	0,80
3	GridSearch	20,40	0,84
	RandomSearch	2,25	0,74
	CMA-ES	10,47	0,79
	TPE	10,34	0,74
4	GridSearch	19,54	0,95
	RandomSearch	2,08	0,91
	CMA-ES	9,85	0,95
	TPE	11,87	0,95
5	GridSearch	20,46	0,91
	RandomSearch	2,21	0,86
	CMA-ES	10,95	0,91
	TPE	10,51	0,86

Сравнение времени работы и точности методов

Полученные доверительные интервалы для скважин

№ 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]

Формирование доверительных интервалов месторождения N

- learning_rate: [0.05, 0.2]
- max_depth: [4, 9]
- min_samples_split: [4, 8]
- n_estimators: [10, 100]

Заключение

- Была изучена задача нахождения оптимальных гиперпараметров для моделей машинного обучения
- Рассмотрены различные методы оптимизации гиперпараметров, такие как GridSearch, RandomSearch, CMA-ES и TPE.
- Проведено сравнение точности и затраченного времени
- Сформированы доверительные интервалы гиперпараметров для моделей скважин и месторождения N