Мищенко И.А., Рубцов Ю.Ф. Прогнозирование объёмного расхода нефти с помощью модели градиентного бустинга // Международный журнал информационных технологий и энергоэффективности. – 2025. – Т. 10 № 12(62) с. 92–99

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.8

ПРОГНОЗИРОВАНИЕ ОБЪЁМНОГО РАСХОДА НЕФТИ С ПОМОЩЬЮ МОДЕЛИ ГРАДИЕНТНОГО БУСТИНГА

[1]Мищенко И.А., Рубцов Ю.Ф.
ФГАОУ ВО "ПЕРМСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ", Пермь, Россия, (614990, Пермский край, город Пермь, Комсомольский пр-кт, д.29), e-mail: [1]rehino@permlink.ru

В статье рассматривается метод прогнозирования объёмного расхода нефти с использованием методов машинного обучения. Актуальность исследования обусловлена потребностью в точном и быстром предсказании расхода нефти по параметрам эксплуатации. Предложен подход, использующий градиентный бустинг решающих деревьев с инженерными и статистическими методами улучшения качества модели. Оценка проводилась с применением кросс-валидации. Метрикой качества служила средняя абсолютная погрешность.

Ключевые слова: Машинное обучение, градиентный бустинг, объёмный расход нефти, прогнозирование, кросс-валидация, регрессионная модель, калибровка.

FORECASTING VOLUMETRIC OIL FLOW USING A GRADIENT BOOSTING MODEL

[1]Mishchenko I. A., Rubtsov Yu. F.
PERM NATIONAL RESEARCH POLYTECHNIC UNIVERSITY, Perm, Russia, (614990, Perm region, Perm, Komsomolskiy pr-kt, 29), e-mail: [1]rehino@permlink.ru

The work considers a method for forecasting the volumetric flow rate of oil based on machine learning techniques. The relevance of the study is driven by the need for accurate and rapid prediction of oil flow using operational parameters. An approach utilizing gradient boosting decision trees is proposed, complemented by engineering and statistical methods to improve model quality. The model was evaluated using cross-validation with mean absolute error as the quality metric.

Keywords: Machine learning, gradient boosting, oil volumetric flow, forecasting, cross-validation, regression model, calibration.

Точное прогнозирование объёмного расхода нефти по параметрам технологического процесса позволяет оптимизировать режимы работы скважин, планировать объём добычи и предотвращать аварийные ситуации.

В настоящее время для оценки расхода нефти широко применяются аналитические формулы и численные гидродинамические модели [1].

К классическим подходам относятся расчёты на основе законов гидравлики и эмпирические корреляции, основанные на законе Дарси и других физических принципах [2].

Однако такие методы требуют тщательного подбора коэффициентов под конкретные условия и могут давать существенные погрешности при изменении режимов работы оборудования. Кроме того, высокоточные численные модели, например, системы

дифференциальных уравнений, решаемые методом конечных элементов или объёмов затратны по времени, и требуют полного набора входных данных, которые не всегда доступны [3].

Для преодоления указанных ограничений было принято решение воспользоваться методами машинного обучения.

Машинное обучение позволяет строить модели напрямую по экспериментальным или производственным данным, минуя необходимость явного задания физико-математической модели процесса.

Алгоритмы ансамблевого обучения, такие как градиентный бустинг деревьев решений, зарекомендовали себя как эффективный инструмент прогнозирования [4].

В отличие от известных подходов, где применяются либо упрощённые эмпирические формулы, либо универсальные модели общего назначения, в работе делается акцент на комбинировании мощного алгоритма бустинга со специальными методами предобработки и калибровки, учитывающими особенности данных (например, различия в диаметрах оборудования).

Исследование включает в себя:

- сбор и подготовку данных измерений расхода нефти и параметров процесса;
- построение модели HistGradientBoostingRegressor с оптимизированными параметрами;
- внедрение логарифмического преобразования целевой переменной и масштабирования признаков;
- введение дополнительных признаков для учёта динамики;
- оценку точности модели методом кросс-валидации;
- анализ влияния каждого из основных факторов на результат и сравнение с существующими подходами.

Для исследования использованы данные замеров объёмного расхода нефти и сопутствующих параметров процесса, полученных на сертифицированном и поверенном расходомере типа «Кориолис» и двух сертифицированных и поверенных датчиках давления.

Набор данных для обучения модели включает ~166 тысяч наблюдений, снятых в различных режимах.

Целевой переменной является объёмный расход нефти (м³/сут).

Значения объёмного расхода варьируются от 1 до 350 м³/сут, то есть охватывают широкий диапазон – от крайне малых до относительно высоких расходов. В качестве базовых признаков в модель включены переменные, приведённые в Таблице 1.

Таблица 1 - Базовые признаки модели

Параметр	**Обозначение**	**Единицы**	**Смысл**
Перепад давления	ΔP	МПа	Разность давления до и после штуцера
Плотность жидкости	ρ	г/см³	Физическое свойство жидкости
Диаметр штуцера	d	мм	Геометрический параметр канала
Температура	T	°C	Влияет на вязкость и плотность
Целевая переменная	Q	м³/сут	Объёмный расход нефти

Для улучшения качества модели применены инженерные и статистические преобразования, отражающие физическую природу течения жидкости:

В формуле (1) отражен дополнительный признак, скоростная зависимость:

$$\sqrt{\Delta P} \tag{1}$$

В формуле (2) отражен дополнительный признак, гидродинамическая функция давления:

$$\sqrt{\frac{\Delta P}{\rho}} \tag{2}$$

В формуле (3) отражен дополнительный признак, площадь сечения канала:

$$d^2 \tag{3}$$

В формуле (4) отражен дополнительный признак, физический прототип формулы расхода:

$$d^2\sqrt{\frac{\Delta P}{\rho}} \tag{4}$$

В формуле (5) отражен дополнительный признак, инверсия плотности:

$$\frac{1}{\sqrt{\rho}} \tag{5}$$

Эти признаки выполняют функции формулы Торричелли (6):

$$Q \propto d^2\sqrt{\frac{\Delta P}{\rho}} \tag{6}$$

Плотность и давление по своему диапазону находились в ограниченных пределах, поэтому для сглаживания распределения без дополнительной нормализации достаточно было логарифмического преобразования по формуле (7):

$$\begin{gathered}\log(1+\Delta P),\\ \log(1+d),\\ \log(1+\rho)\end{gathered} \tag{7}$$

Так стабилизируется масштаб и зависимость становится более линейной для бустинга.

В частности, целевая переменная Q также заменена на log(Q), при обучении модели. Переход к логарифму расхода позволил уменьшить влияние асимметричного распределения Q и сделать задачу более удобной для регрессии.

Температура вместо перехода в логарифмическое пространство была нормализована по формуле (8):

$$\mathrm{T}_z = \frac{T-\mu_T}{\sigma_T} \tag{8}$$

, где μ_T и σ_T — среднее и стандартное отклонение в обучающем фолде.

В итоге модель использует 13 признаков, объединяющих физические и статистические описания входных данных.

Перед обучением модели проведена очистка данных. Удалены некорректные записи, в частности отрицательные значения расхода и других параметров. Также из набора исключены явные выбросы, не соответствующие физически возможным режимам, например, случаи нулевого или экстремально большого перепада давления при значительном расходе.

В качестве алгоритма прогнозирования выбран градиентный бустинг решающих деревьев, а именно HistGradientBoostingRegressor из библиотеки scikit-learn [5].

Данный алгоритм строит ансамбль из M неглубоких деревьев решений (9), каждое последующее дерево обучается на ошибках предыдущих, таким образом постепенно уменьшая ошибку ансамбля.

$$\hat{y} = \sum_{m=1}^{M} \eta * \mathrm{h_m}(x) \quad (9)$$

В формуле (9):

$\mathrm{h_m}(x)$ – это m-ное дерево, аппроксимирующее отрицательный градиент ошибки на предыдущем шаге,

η – это шаг при обучении,

M – число итераций бустинга.

Особенностью HistGradientBoosting (HGB) является эффективный алгоритм бустинга на основе гистограмм, что ускоряет обучение на больших выборках за счёт биннинга непрерывных признаков [6].

Для нашей задачи HGBRegressor был настроен с использованием функции потерь на основе абсолютной погрешности (10) (mean absolute error, MAE).

$$\mathrm{L(y}, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{y_i} - \hat{y_i}) \quad (10)$$

Выбор MAE (вместо стандартной квадратичной ошибки) продиктован тем, что абсолютная погрешность менее чувствительна к выбросам и даёт более устойчивые медианные оценки, что важно при наличии редких аномально высоких или низких значений расхода [7].

Для оценки качества применялась 5-кратная перекрёстная проверка (11):

$$D = \bigcup_{k=1}^{5} (D_{train}^{k}, D_{val}^{k}) \quad (11)$$

Метрики усреднялись по всем фолдам. Использовались следующие показатели:

$$MAE = \frac{1}{n}\sum |y - \hat{y}| \quad (12)$$

$$MAPE = \frac{100}{n}\sum \frac{|y - \hat{y}|}{|y|} \quad (13)$$

$$SMAPE = \frac{200}{n}\sum \frac{|y - \hat{y}|}{|y| + |\hat{y}|} \quad (14)$$

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \hat{y})^2} \quad (15)$$

Для оценки равномерности погрешности также вычисляется доля прогнозов в допуске ±5% и ±10%.

При обучении модели была использована схема взвешивания наблюдений по относительной погрешности, чтобы обеспечить равную значимость относительных отклонений [8].

Для малых по величине расходов даже небольшое абсолютное отклонение может быть критичным в процентном отношении, тогда как для больших расходов такая же абсолютная погрешность может составлять доли процента.

Важно отметить, что для предотвращения переобучения и обеспечения физической осмысленности результатов рассматривалась возможность введения монотонных ограничений [9]. Однако в конечном итоге было принято не вводить монотонные ограничения, поскольку данные содержат сложные взаимосвязи.

После первоначального обучения была проведена дополнительная калибровка прогнозов, призванная устранить систематические смещения в зависимости от диаметра штуцера.

Анализ разностей между прогнозом модели и фактическим значением показал, что даже при общем небольшом уровне погрешности модель имеет тенденцию к небольшому недопредсказанию или перепредсказанию расхода для определённых диаметров.

Чтобы учесть эту особенность, была применена постобучающая калибровка по диаметру двумя методами: линейной регрессией (16) и изотонической регрессией (17), [10].

$$Q' = a_d + b_d \hat{Q} \tag{16}$$

где a_d и b_d подбираются методом наименьших квадратов по каждому диаметру.

$$Q'' = Iso_d(Q') \tag{17}$$

где Iso_d монотонная аппроксимация зависимости истинного расхода от предсказанного для данного диаметра.

Обе схемы калибровки сравнивались между собой.

Изотоническая калибровка лучше устраняет смещения: средняя погрешность после неё оказалась ниже, особенно на краях диапазона.

Линейная калибровка также улучшила точность по сравнению с некалиброванной моделью, но недостаточно корректировала нелинейные эффекты.

Таблица 2 - Результаты кросс-валидации

Метрика	**Значение**
RMSE	1.995
MAE	1.223
R^2	0.899629
MAPE	2.260%
SMAPE	2.158%
WMAPE	1.121%
Покрытие ±5%	91.87%
Покрытие ±10%	96.81%

После обучения и кросс-валидации модель градиентного бустинга продемонстрировала высокое качество прогнозирования. Среднее значение MAE по итогам 5-фолдовой кросс-валидации составило 1.223 м³/сут.

Такой уровень погрешности свидетельствует о том, что предложенный алгоритм в целом способен с высокой точностью воспроизводить реальные значения расхода нефти.

Также следует отметить низкое стандартное отклонение MAE между фолдами, что указывает на стабильность модели и её обобщающую способность на разных подмножествах данных.

Итоговое сравнение предсказанного объёмного расхода нефти с настоящим объёмным расходом нефти представлено на Рисунке 1.

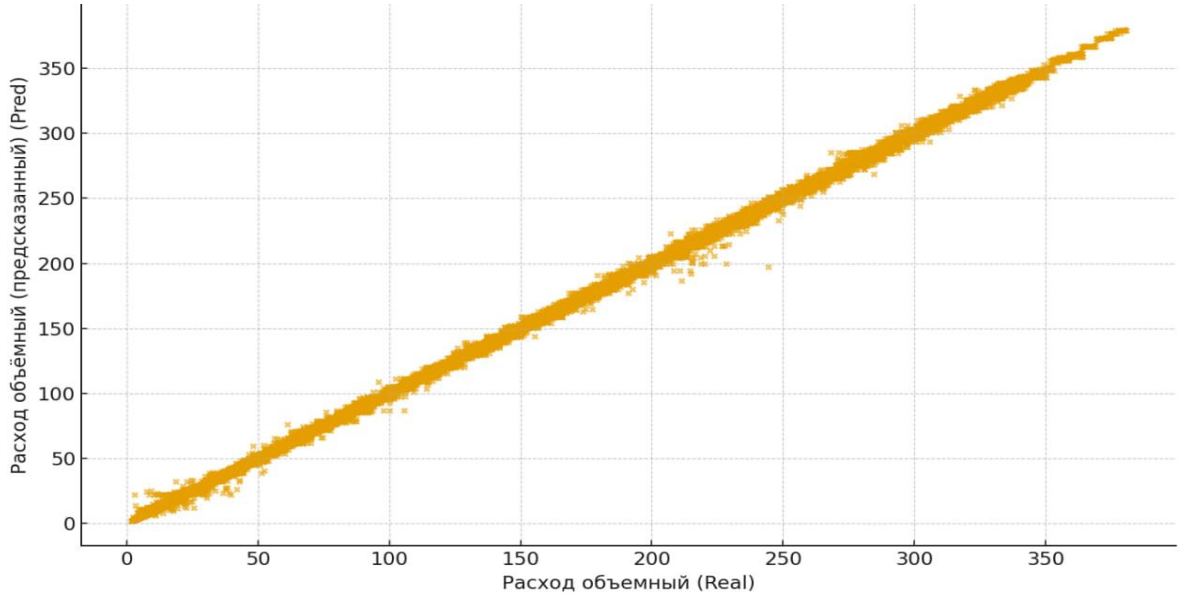

Рис унок1 - Сравнение предсказанного и реального объёмного расхода нефти

Распределение относительной погрешности в зависимости от расхода показано на Рисунке 2.

Наибольшая погрешность возникает на низких расходах, но на графике видно, что погрешность равномерно распределена и какие-либо выраженные систематическая зависимости отсутствуют.

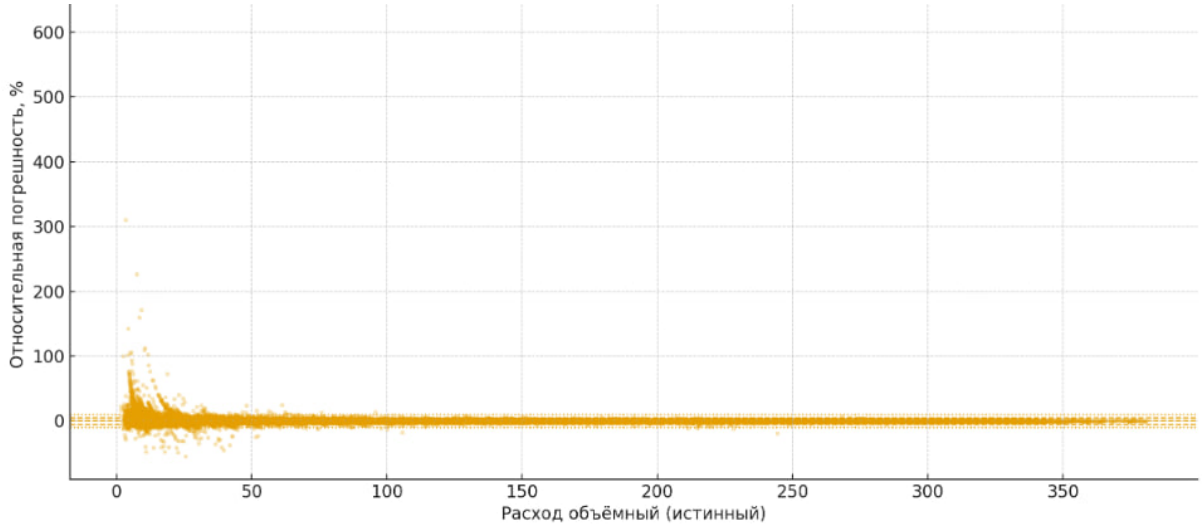

Рисунок 2 - Зависимость погрешности от величины реального расхода нефти

В ходе исследования было проведено прогнозирование объёмного расхода нефти на основе современной модели машинного обучения – градиентного бустинга решающих деревьев.

Разработанный подход позволил учесть одновременно влияние основных технологических параметров (давление, диаметр штуцера, плотность и температура жидкости) и выявить нелинейные зависимости, неочевидные при использовании традиционных формул.

Использование логарифмического преобразования целевой переменной и взвешивание ошибок по относительной значимости обеспечили высокую точность прогноза по всему диапазону расходов.

Средняя относительная погрешность модели составила всего 2.5%, что свидетельствует о высоком качестве предсказаний.

В результате разработанная модель сочетает в себе преимущества градиентного бустинга и аддитивной пост-обучающей поправки.

Полученные зависимости соответствуют физическому смыслу, что подтверждает корректность подхода, а погрешность находится на приемлемом уровне для практических расчётов.

Практическая ценность работы состоит в том, что предложенная модель может быть внедрена в систему оперативного мониторинга и управления, выдавая прогноз расхода нефти в реальном времени на основе текущих показателей давления, температуры и плотности.

Высокая скорость работы алгоритма градиентного бустинга и его точность открывают возможности для прогнозирования аварийных ситуаций и оптимизации режима работы нефтяных скважин.

Список литературы

1. Пискунов С. А., Давуди Ш. Прогнозирование дебита горизонтальных скважин с применением модели машинного обучения // Известия Томского политехнического университета. Инжиниринг георесурсов. – 2024. – Т. 335. – № 5. – С. 107–117.
2. Friedman J. H. Greedy function approximation: A gradient boosting machine // Annals of Statistics. – 2001. – Vol. 29. – No. 5. – pp. 1189–1232.
3. Natekin A., Knoll A. Gradient boosting machines, a tutorial // Frontiers in Neurorobotics. – 2013. – Vol. 7. – Article 21.
4. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – pp. 2825–2830.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. 2nd ed. – New York: Springer, 2009. – p.745
6. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’16). – New York: ACM, 2016. – pp. 785–794.
7. Ke G., Meng Q., Finley T., et al. LightGBM: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems 30 (NIPS 2017). – 2017. – pp. 3146–3154.
8. Bahrami S., Rashidi F., Shokrollahi A., et al. Estimation of flow rates of individual phases in oil–gas–water flows using machine learning techniques // Flow Measurement and Instrumentation. – 2019. – Vol. 66. – pp. 28–36.
9. Goes L. C. S., Gildin E., Santos S. V. Virtual flow metering of oil wells for a pre-salt field // Journal of Petroleum Science and Engineering. – 2021. – Vol. 203. – Article 108586.

10. Alkhammash E. H., Qasem S. N., et al. An optimized gradient boosting model by genetic algorithm for forecasting crude oil production // Energies. – 2022. – Vol. 15. – No. 17. – p. 6416.

References

1. Piskunov S. A., Davudi Sh. Forecasting the production rate of horizontal wells using a machine learning model // Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering. 2024. T. 335. No. 5. pp. 107–117.
2. Friedman J. H. Greedy function approximation: a gradient boosting machine // Annals of Statistics. 2001. Vol. 29. No. 5. pp. 1189–1232.
3. Natekin A., Knoll A. Gradient boosting machines, a tutorial // Frontiers in Neurorobotics. 2013. Vol. 7. Art. 21.
4. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. pp. 2825–2830.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. p.745
6. Chen T., Guestrin C. XGBoost: a scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). New York, 2016. pp. 785–794.
7. Ke G., Meng Q., Finley T., et al. LightGBM: a highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems. 2017. Vol. 30. pp. 3146–3154.
8. Bahrami S., Rashidi F., Shokrollahi A., et al. Estimation of flow rates of individual phases in oil–gas–water flows using machine learning techniques // Flow Measurement and Instrumentation. 2019. Vol. 66. pp. 28–36.
9. Goes L. C. S., Gildin E., Santos S. V. Virtual flow metering of oil wells for a pre-salt field // Journal of Petroleum Science and Engineering. 2021. Vol. 203. Art. 108586.
10. Alkhammash E. H., Qasem S. N., et al. An optimized gradient boosting model by genetic algorithm for forecasting crude oil production // Energies. 2022. Vol. 15. No. 17. p. 6416.