УДК 004.896

Разработка объяснимых моделей для виртуального анализа содержания серы в нефтепродуктах

Кобаченко Александра Андреевна kobachenkoo@mail.ru
Мурашов Михаил Владимирович murashov@bmstu.ru
Герасименко Илья Владимирович ilj4ger@yandex.ru

МГТУ им. Н.Э. Баумана, Москва, Россия

В работе рассматриваются объяснимые, или интерпретируемые, методы построения виртуальных анализаторов для прогнозирования содержания серы в бутан-бутиленовой фракции. Проведен сравнительный анализ линейной регрессии с весами, деревьев решений различной глубины, случайного леса и градиентного бустинга. Выполнено исследование зависимости точности прогноза от глубины деревьев и проанализированы важности признаков, что позволило установить физико-технологическую интерпретацию полученных зависимостей. Результаты показали, что ансамблевые модели на основе деревьев средней глубины обеспечивают наилучший баланс между точностью прогнозирования и объяснимостью, что делает их перспективными для промышленного внедрения.

Ключевые слова: *виртуальный анализатор, машинное обучение, контроль качества, нефтепереработка, прогнозирование качества*

Введение. Контроль содержания серы в углеводородных фракциях является одной из ключевых задач нефтепереработки. Превышение нормативных значений ведет к деградации катализаторов, ускоренной коррозии оборудования и нарушению экологических регламентов [1]. Традиционные методы контроля, основанные на лабораторных и поточных анализаторах, обеспечивают высокую точность, однако обладают рядом ограничений: значительными временными задержками, высокой стоимостью обслуживания и отсутствием непрерывности измерений.

Виртуальные анализаторы (ВА), основанные на алгоритмах машинного обучения, представляют собой математические модели, способные в реальном времени прогнозировать содержание примесей по косвенным технологическим параметрам (температура, давление, расходы реагентов и др.) [2]. Важнейшим требованием для промышленного применения является высокая точность прогноза и объяснимость моделей, позволяющая технологам интерпретировать влияние параметров процесса на качество продукции.

Цель настоящего исследования заключается в разработке и сравнительном анализе интерпретируемых моделей виртуального анализа содержания серы с использованием регрессионных и ансамблевых методов.

Математическая постановка задачи. Пусть имеется набор наблюдений

$$(X_i, y_i)\}_{i=1}^{n}, X_i = [x_{i1}, x_{i2}, x_{i3}],$$

где X_i — вектор технологических параметров; y_i — содержание серы в образце, определенное лабораторным методом.

Требуется построить модель $\hat{y} = f(X)$, минимизирующую функцию ошибки

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| = \text{MAE}.$$

Качество аппроксимации дополнительно оценивается коэффициентом детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Предлагаемое решение. В рамках настоящего исследования рассмотрены несколько классов интерпретируемых моделей: линейная регрессия с весами, деревья решений, ансамблевые методы (случайный лес и градиентный бустинг). Входными параметрами модели выступают три технологических параметра: x_1 — расход щелочного раствора на входе; x_2 — давление в регенераторе щелочи; x_3 — расход гидроочищенного бензина.

Первая модель реализует базовый метод, чаще всего использующийся на реальных производствах — метод взвешенных наименьших квадратов (WLS), где каждой точке назначается вес w_i:

$$S_w(\beta) = \sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

Таким образом, наблюдения с высокой дисперсией или выбросами оказывают меньшее влияние на итоговую модель [3].

Достоинства: высокая интерпретируемость коэффициентов, простота реализации.

Недостатки: невозможность учета нелинейных зависимостей и взаимодействий признаков.

Согласно результатам обучения на исторических данных, итоговое регрессионное уравнение для прогнозирования содержания серы имеет вид

$$y = 725{,}78 - 58{,}23 \cdot x_1 - 454{,}03 \cdot x_2 - 0{,}13 \cdot x_3.$$

На рис. 1 представлен результат работы модели взвешенных наименьших квадратов.

Из рис. 1 можно сделать вывод, что модель WLS продемонстрировала ограниченную точность: средняя абсолютная ошибка составила 16,71 ppm, коэффициент детерминации R2 = 0,55. Это значит, что линейная аппроксимация действительно не способна в полной мере отразить сложные нелинейные зависимости, необходимо использование методов машинного обучения (МО).

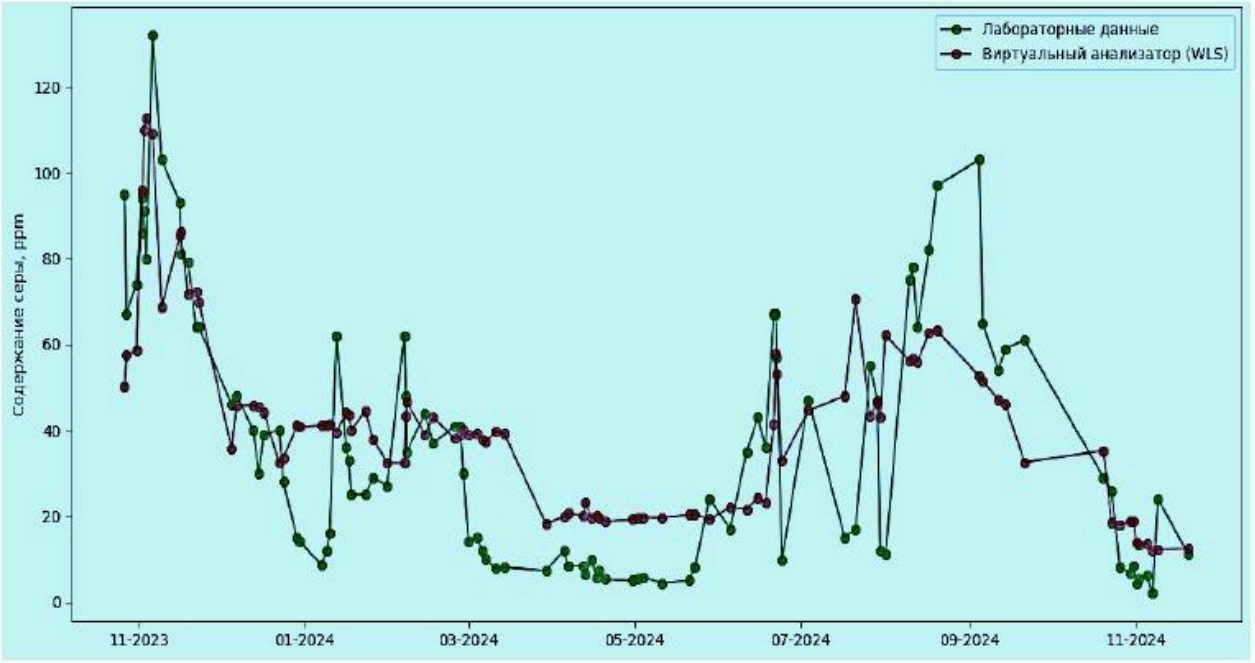

Рис. 1. Результат работы модели WLS

Рассмотрим модель «дерево решений». На каждом шаге выбирается предикат вида: $x_j < t$, который минимизирует функционал неоднородности в выборках.

Для регрессионных задач критерием качества является минимизация внутриклассовой дисперсии

$$Q = \sum_{i \in R_1} \left(y_i - \widehat{y_{R_1}}\right)^2 + \sum_{i \in R_2} \left(y_i - \widehat{y_{R_2}}\right)^2,$$

где R_1, R_2 — разбиения выборки [4, 5].

Построенное дерево имеет глубину *d*, которая регулирует баланс между точностью и объяснимостью. Достоинства: наглядная интерпретация в виде правил «если — то», выявление пороговых зависимостей. Недостатки: нестабильность к изменению данных, снижение объяснимости при большой глубине.

Проведено исследование зависимости точности от глубины дерева (табл. 1).

Таблица 1

Зависимость качества модели от глубины дерева

Глубина	MAE(ppm)	R^2
4	11,5	0,74
5	10,9	0,76
6	10,7	0,77
7	10,6	0,78

Установлено, что оптимальной является глубина 5–6 уровней: при дальнейшем увеличении прирост точности несущественен, при этом снижается интерпретируемость.

Ансамбль из M деревьев строится по схеме

$$\widehat{y_{RF}}(X) = \frac{1}{M}\sum_{m=1}^{M} T_m(X),$$

где T_m — отдельное дерево решений.

Важности признаков: $\mathrm{Imp}(x_1) = 0{,}47$, $\mathrm{Imp}(x_2) = 0{,}41$, $\mathrm{Imp}(x_3) = 0{,}12$.

Пусть $\hat{y}_{WLS}$ — прогноз, полученный с использованием модели WLS, $\hat{y}_{RF}$ — прогноз, полученный с помощью модели Random Forest, $\hat{y}_{ens}$ — итоговый ансамблевый прогноз. Тогда итоговый прогноз ансамбля вычисляется по формуле

$\hat{y}_{ens} = \alpha \cdot \hat{y}_{WLS} + (1 - \alpha) \cdot \hat{y}_{RF}$.

На рис. 2 представлен результат работы ансамблевой модели.

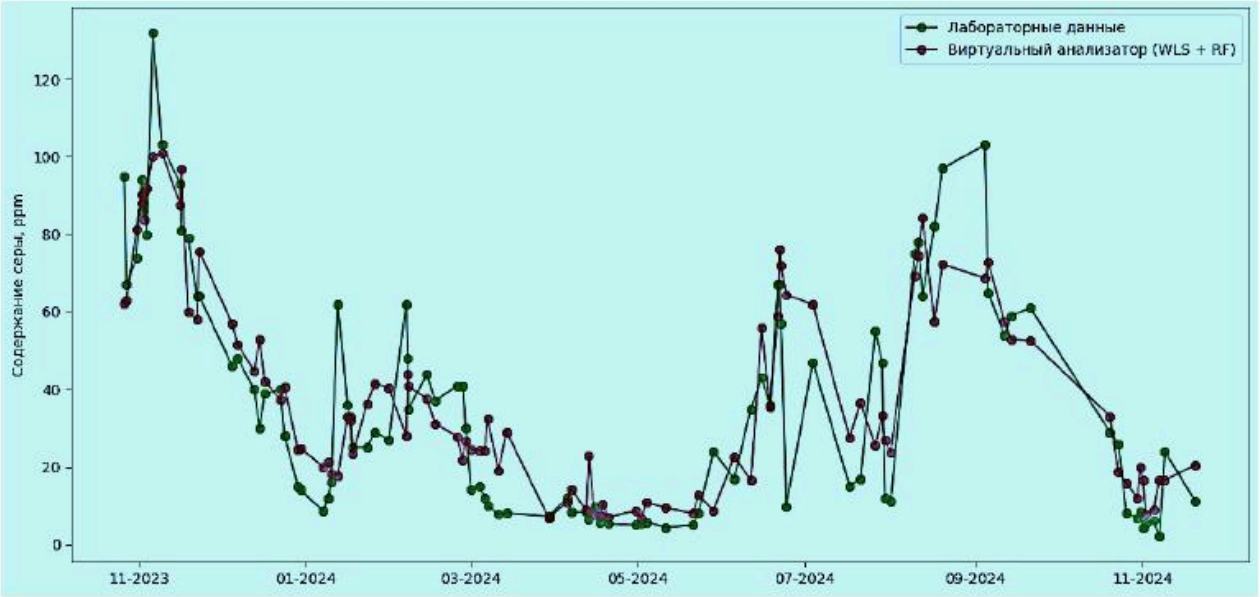

Рис. 2. Результат работы ансамблевой модели

Из рис. 2 видно, что ансамбль продемонстрировал заметное улучшение по сравнению с WLS: средняя абсолютная ошибка составила 10,77 ppm, R^2 = 0,77.

Метод градиентного бустинга представляет собой последовательное построение ансамбля деревьев. Модель формируется последовательным добавлением деревьев:

$$\widehat{y^{(t)}} = \widehat{y^{(t-1)}} + \eta \cdot T_t(X),$$

где η — коэффициент обучения.

Результаты: MAE = 10,1 ppm, $R^2 = 0{,}80$. Важности признаков распределились следующим образом: $\mathrm{Imp}(x_1) = 0{,}39$, $\mathrm{Imp}(x_3) = 0{,}17$.

Особенностью XGBoost является введение регуляризации, которая накладывает ограничения на сложность деревьев:

$$\Omega(T) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2,$$

где J — число листьев; w_j — веса предсказаний в листьях; γ, λ — коэффициенты регуляризации.

Достоинства: высокая точность, учет сложных нелинейных зависимостей, возможность интерпретации через методы SHAP и feature importance.

Недостатки: высокая вычислительная сложность, необходимость настройки гиперпараметров.

Выводы. Линейная модель WLS обладает высокой интерпретируемостью, но недостаточной точностью ($R^2 = 0{,}55$). Деревья решений позволяют выявлять нелинейные зависимости и пороговые эффекты, однако при чрезмерной глубине теряют объяснимость. Случайный лес обеспечивает хорошую устойчивость и точность при сохранении возможности анализа важности признаков. Градиентный бустинг демонстрирует наилучшее качество прогнозирования, оставаясь частично объяснимым за счет методов анализа влияния признаков (feature importance, SHAP-значения).

Литература

[1] Капустин В.М., Рудин М.Г. *Химия и технология нефтепереработки*. Москва, Химия, 2013, 496 с.

[2] Шевлягина С.А., Торгашов А.Ю. *Методы построения виртуальных анализаторов в системах усовершенствованного управления технологическими процессами*. Владивосток, Изд-во Дальневост. федерал. ун-та, 2024, 1 CD, 86 с.

[3] Кобаченко А.А., Паротькина М.А., Егоров С.В., Волынчиков М.А. Предобработка данных, поступающих на виртуальный анализатор нефтеперерабатывающего предприятия. *Робототехника и искусственный интеллект. XVI Всерос. науч.-техн. конф. с междунар. уч.: матер*. Красноярск, ЛИТЕРА-принт, 2024, с. 130–133.

[4] Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* California, Springer, 2009, 746 p.

[5] Rožanec J, Trajkova E., Jinzhi L., Sarantinoudis N., Arampatzis G. Eirinakis P., Mourtos I., Onat M., Yılmaz D., Košmerlj A., Kenda K., Fortuna B., Mladenić D. Cyber-Physical LPG Debutanizer Distillation Columns: Machine-Learning-Based Soft Sensors for Product Quality Monitoring. *Applied Sciences*, 2021, vol. 11, art. no. 11790. https://doi.org/10.3390/app112411790

Development of explainable models for virtual analysis of sulfur content in petroleum products

Kobachenko Aleksandra Andreevna kobachenkoo@mail.ru
Murashov Mikhail Vladimirovich murashov@bmstu.ru
Gerasimenko Ilya Vladimirovich ilj4ger@yandex.ru

BMSTU, Moscow, Russia

The paper considers explicable, or interpretable, methods for constructing virtual analyzers for predicting the sulfur content in the butane-butylene fraction. A comparative analysis of linear regression with weights, decision trees of various depths, random forest and gradient boosting is carried out. The dependence of the forecast accuracy on the depth of the trees was studied and the importance of the features was analyzed, which made it possible to establish a physico-technological interpretation of the obtained dependencies.

Keywords: *softt sensor, machine learning, quality control, oil refining, quality forecasting*