УДК 504.3.054:004.852

Коробейников Илья Алексеевич, магистрант
(ДВФУ, г. Владивосток)
Горошко Сергей Дмитриевич, магистрант
(ДВФУ, г. Владивосток)
Голубева Алина Константиновна, магистрант
(ДВФУ, г. Владивосток)
Синодзаки Анастасия, магистрант
(ДВФУ, г. Владивосток)
Абрецова Елена Сергеевна, магистрант
(КузГТУ, г. Кемерово)

Ilya A. Korobeinikov, undergraduate
(FEFU, Vladivostok)
Sergey D. Goroshko, undergraduate
(FEFU, Vladivostok)
Alina K. Golubeva, undergraduate
(FEFU, Vladivostok)
Anastasia Shinozaki, undergraduate
(FEFU, Vladivostok)
Elena S. Abretsova, undergraduate
(KuzSTU, Kemerovo)

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ ГРАДИЕНТНОГО БУСТИНГА ДЛЯ КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ КОНЦЕНТРАЦИИ ДИОКСИДА АЗОТА В АТМОСФЕРНОМ ВОЗДУХЕ

EVALUATING GRADIENT BOOSTING ALGORITHMS FOR SHORT-TERM FORECASTING OF AMBIENT NITROGEN DIOXIDE CONCENTRATIONS

Аннотация: На данных UCI Air Quality сопоставлены XGBoost, Random Forest и Persistence. На горизонте 1 ч лучшим оказался Persistence; на 3 и 6 ч XGBoost с календарными признаками снизил MAE на 19,56 и 32,07%. Добавление сенсоров не улучшило результат.

Abstract: On UCI Air Quality data, XGBoost and Random Forest were compared with persistence. Persistence performed best at 1 h. At 3 and 6 h, XGBoost with calendar features reduced MAE by 19.56% and 32.07%. Adding sensor measurements did not improve performance.

Введение

Краткосрочный прогноз NO_2 дополняет мониторинг качества воздуха. Преимущество сложной модели над переносом последнего измерения необходимо проверять на более позднем периоде. Цель работы — оценить влияние горизонта и состава признаков на качество XGBoost в сравнении с Random Forest и Persistence. Проверяется гипотеза о пользе календаря и текущих измерений среды. Исследование выполнено на итальянских данных UCI [1], а не на данных Кузбасса.

Материалы и методы

Использован оригинальный AirQualityUCI.csv из UCI Air Quality [1], загруженный 24.10.2026. В набор входят эталонные концентрации и отклики пяти металлооксидных сенсоров; связанная работа рассматривает их полевую калибровку [2]. CSV прочитан с разделителем «;» и десятичной запятой, −200 заменено на пропуск. Удалены 114 пустых строк и два пустых столбца.

Таблица 1 — Характеристики исходных данных

Показатель	Значение
Почасовые записи	9357
Период CSV	10.03.2004–04.04.2005
Пропуски NO_2	1642 (17,55%)
Пропуски каждого из T, RH, AH и пяти сенсоров	366 (3,91%)
Дубликаты / отсутствующие часы	0 / 0



Рисунок 1 — Эталонный ряд NO_2; пропуски показаны разрывами

Фактический период файла отличается от краткого описания UCI, заканчивающегося февралем 2005 г. Временные метки использованы без

предположений о часовом поясе. Целевая концентрация измеряется в мкг/м³.

Постановка прогноза и протокол

Для каждого часа t отдельная модель прогнозирует концентрацию через h часов. Текущее $NO_2(t)$ считается известным; все входные наблюдения имеют временные метки не позже t. Рекурсивная подстановка собственных прогнозов отсутствует.

$$\hat{C}(t+h) = f_h(X_t), \quad h \in \{1,3,6\}. \tag{1}$$

F1 включает NO_2 в t, t−1, t−3, t−6 и t−24 (5 признаков). F2 добавляет sin/cos часа суток и дня недели (9). F3 добавляет T, RH, AH и PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3) (17). Прочие эталонные GT-показания исключены.

Сначала проверена регулярная почасовая сетка, затем сформированы лаги и цели строго t+h. Использованы только строки с наблюдаемыми $NO_2(t)$ и $NO_2(t+h)$; цели не интерполировались. Входные пропуски заменены медианами обучающей части, одинаково для обеих ML-моделей.

Общая календарная шкала разделена 60/20/20% по времени цели: validation начинается 30.10.2004 в 16:00, test — 16.01.2005 в 15:00. Дополнительно исключены прогнозы validation/test, выпущенные до начала соответствующего периода. Внутри горизонта все модели и группы имеют одинаковые тестовые часы.

Таблица 2 — Число допустимых примеров

h, ч	Train	Validation	Test
1	4019	1593	1756
3	3978	1588	1753
6	3923	1579	1752

Persistence переносит последнее измерение на будущий момент. Random Forest [3, 5] использует 250 деревьев; проверены пары (глубина, минимум объектов в листе): (10, 2), (без ограничения, 2), (10, 5). Для XGBoost [4, 6] проверены 6 конфигураций: 200–400 деревьев, глубина 2–5, learning_rate 0,03–0,10, subsample и colsample_bytree 0,8–1,0, reg_lambda 1–10.

Всего обучен 81 кандидат. Выбор — по минимальному validation MAE отдельно для каждой модели, группы и горизонта. Обучение только на train, без финального переобучения. CPU, один поток, seed=42; Python 3.12.10, scikit-learn 1.9.1, XGBoost 3.4.1. Тест не использован для настройки.

Метрики и результаты

Использованы MAE и RMSE в мкг/м³, R^2 и относительное снижение MAE к Persistence. Отрицательные R^2 и улучшения сохранены.

$$MAE = \frac{\Sigma_i |y_i - \hat{y}_i|}{n}, \quad RMSE = \sqrt{\left(\frac{\Sigma_i (y_i - \hat{y}_i)^2}{n}\right)}. \tag{2}$$

$$R^2 = 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y})^2}, \quad \Delta MAE = 100 \frac{MAE_p - MAE_m}{MAE_p}. \tag{3}$$

Здесь n — число примеров, y_i — факт, $\hat{y}_i$ — прогноз, $\bar{y}$ — среднее фактических значений; p — Persistence, m — модель.

Таблица 3 — Полные тестовые оценки

h, ч	Гр.	Модель	MAE	RMSE	R^2	ΔMAE, %
1	—	P	19,798	26,735	0,745	0,00
1	F1	RF	22,493	31,964	0,635	-13,61
1	F1	XGB	22,278	31,464	0,647	-12,53
1	F2	RF	22,080	31,959	0,635	-11,53
1	F2	XGB	20,955	30,138	0,676	-5,84
1	F3	RF	23,122	33,219	0,606	-16,79
1	F3	XGB	24,223	34,480	0,576	-22,35
3	—	P	43,398	55,290	-0,074	0,00
3	F1	RF	43,334	56,571	-0,125	0,15
3	F1	XGB	46,175	60,383	-0,281	-6,40
3	F2	RF	36,211	48,765	0,164	16,56
3	F2	XGB	34,908	46,968	0,225	19,56
3	F3	RF	38,056	51,059	0,084	12,31
3	F3	XGB	36,058	48,481	0,174	16,91
6	—	P	56,291	69,564	-0,730	0,00
6	F1	RF	56,167	72,166	-0,862	0,22
6	F1	XGB	54,978	71,104	-0,808	2,33
6	F2	RF	38,254	50,079	0,103	32,04
6	F2	XGB	38,237	50,186	0,099	32,07
6	F3	RF	40,382	52,468	0,016	28,26
6	F3	XGB	39,960	52,306	0,022	29,01

P — Persistence; RF — Random Forest; XGB — XGBoost. Persistence одинаков для F1–F3 и показан один раз. N_test: 1756, 1753 и 1752 при h = 1, 3 и 6 ч соответственно. Округление выполнено только для представления.

Влияние горизонта и состава признаков

На 1 ч Persistence лучше всех выбранных ML-моделей: MAE = 19,798. XGBoost F2 дает 20,955, то есть ошибка выше на 5,84%. На 3 и 6 ч XGBoost F2 снижает MAE к Persistence на 19,56 и 32,07% соответственно (рис. 2). Это относительное снижение ошибки, а не процент увеличения «точности».

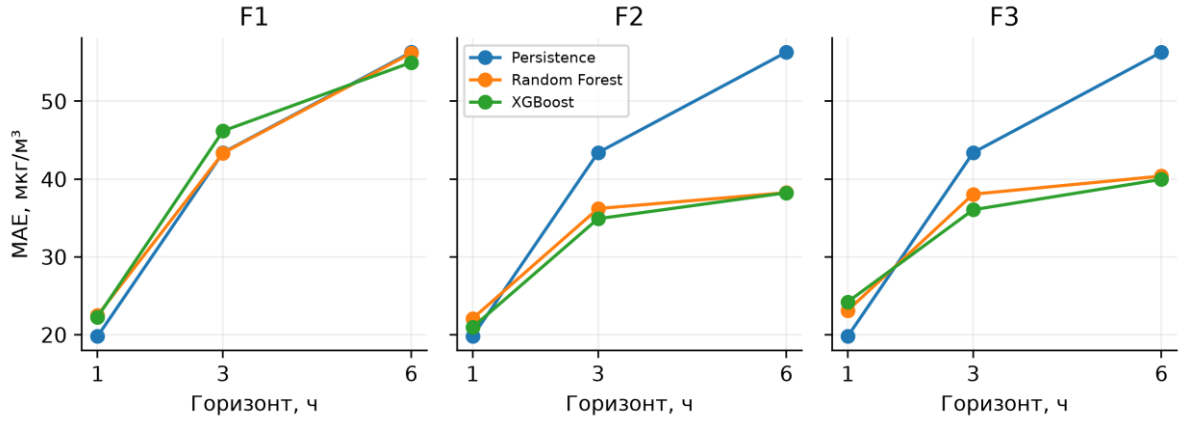

Рисунок 2 — Тестовый MAE для трех групп признаков

Календарь снижает MAE XGBoost относительно F1 на всех горизонтах. F3, напротив, уступает F2: 24,223; 36,058; 39,960 против 20,955; 34,908; 38,237 мкг/м³. У RF направление то же. Польза добавления сенсоров и параметров среды не подтверждена. Вывод относится к выбранному разбиению и ограниченному поиску параметров.

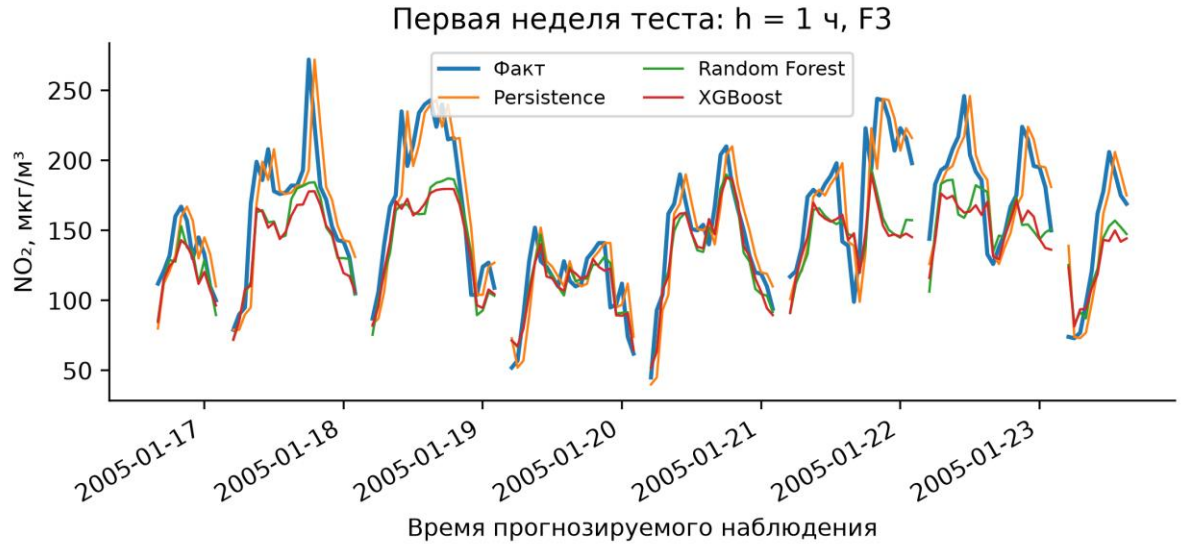

Рисунок 3 — Первая неделя теста: h = 1 ч, F3

На 6 ч MAE XGBoost F2 (38,237) и RF F2 (38,254) практически совпадают; RMSE у RF немного ниже. Статистически значимое превосходство XGBoost над RF не установлено.

Объяснение модели и пропуски входов

Для заранее выбранного h=1/F3 рассчитаны точные TreeSHAP-вклады XGBoost [7] на всех 1756 тестовых строках. Проверена аддитивность вкладов. Наиболее велик средний абсолютный вклад $NO_2(t)$ — 35,320 мкг/м³, затем cos часа — 6,724 и PT08.S2(NMHC) — 3,503. Это объяснение модели, а не доказательство причинного влияния факторов.

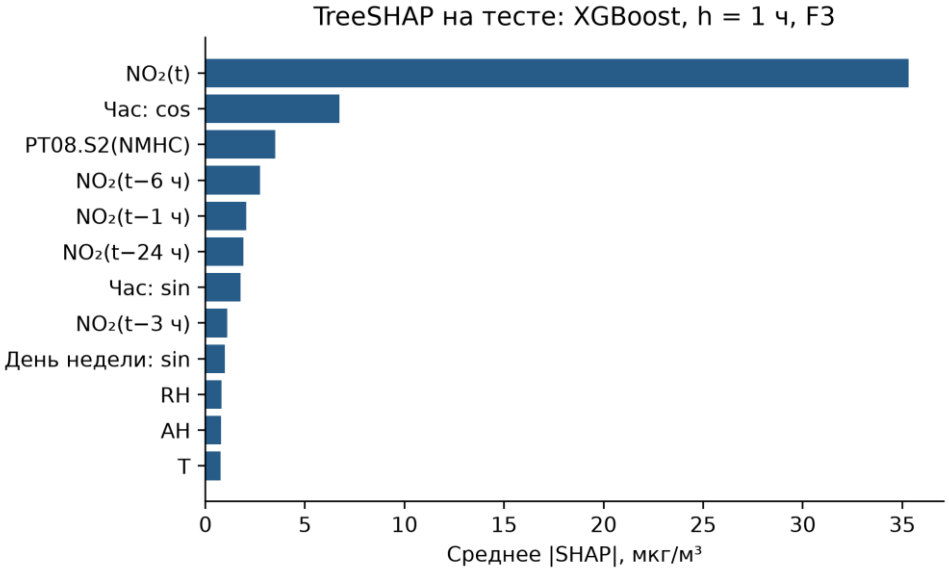

Рисунок 4 — Средние абсолютные TreeSHAP-вклады

Стресс-тест выполнен для h=1/F3: маскировались ячейки сформированной матрицы, кроме $NO_2(t)$ и календаря. Маски вложены, одинаковы для моделей, seed=42. Истинные цели неизменны, модели не переобучались. Это не физический отказ телеметрии: лаги скрываются независимо. Persistence определен на тех же 1756 строках.

Таблица 4 — MAE при дополнительных пропусках, мкг/м³

Уровень, %	Persistence	Random Forest	XGBoost
0	19,798	23,122	24,223
5	19,798	23,290	24,408
15	19,798	23,702	24,747
30	19,798	24,280	25,220

Уровень — вероятность маскирования, а не точная доля новых пропусков. Скрыто 0, 1020, 2990 и 6020 ранее наблюдаемых ячеек. Использована одна реализация маски; результат не описывает длительные сбои.

Обсуждение

Преимущество Persistence на 1 ч согласуется с полезностью последнего измерения на ближайшем горизонте. При увеличении горизонта календарь помогает описать повторяющиеся временные закономерности. Дополнительные измерения среды не дали выигрыша. Возможные объяснения — дрейф сенсоров, нестационарность и коррелированность откликов; настоящий эксперимент не устанавливает причину ухудшения.

Источник представляет один устаревший локализованный итальянский ряд. Не хватает ветра, давления и высоты слоя перемешивания. Температура и влажность не заменяют эти факторы. Отбор часов с доступными $NO_2(t)$ и $NO_2(t+h)$ может смещать оценку, если пропуски связаны с режимами загрязнения. Медианная импутация не учитывает временную структуру пропусков.

Один holdout, один seed и ограниченная сетка не дают универсального рейтинга алгоритмов. Доверительные интервалы не оценивались. Между горизонтами наборы часов немного отличаются; внутри горизонта они одинаковы для моделей и групп.

Фиксированные модели обучены на раннем периоде и не обновляются на тесте. При каждом новом прогнозе доступны уже поступившие наблюдения, включая $NO_2(t)$; это не прогноз всего тестового периода из одной начальной точки. При задержке публикации эталонного NO_2 постановку необходимо изменить.

Переносимость на Кузбасс не доказана. Для применения необходимы местные данные и повторная временная валидация. Результаты не обосновывают причинные экологические выводы или готовность системы предупреждения опасных эпизодов.

Воспроизводимость

Сохранены исходные данные и SHA-256, протокол, параметры, прогнозы и маски пропусков. Два полных запуска дали точное совпадение численных результатов. Пройдены 14 тестов временного выравнивания, причинности признаков, импутации только на train и сопоставимости тестовых меток.

Заключение

На изученном разбиении для 1 ч предпочтителен Persistence. На 3 и 6 ч XGBoost с историей NO_2 и календарем снижает MAE на 19,56 и 32,07%. Добавление текущих сенсорных и температурно-влажностных измерений ухудшило результат относительно F2. На 6 ч XGBoost и RF F2 близки. Пропуски входной матрицы ухудшили обе ML-модели при сохраненном $NO_2(t)$. Выбор алгоритма должен учитывать горизонт и проверяться на независимом более позднем периоде.

Список литературы

1. Vito S. Air Quality [Dataset]. UCI Machine Learning Repository, 2008. DOI: 10.24432/C59K5F.
2. De Vito S., Massera E., Piga M., Martinotto L., Di Francia G. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario // Sensors and Actuators B: Chemical. 2008. Vol. 129, no. 2. P. 750–757. DOI: 10.1016/j.snb.2007.09.060.
3. Breiman L. Random Forests. University of California, Berkeley, 2001. URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (дата обращения: 24.10.2026).
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016. arXiv:1603.02754. URL: https://arxiv.org/abs/1603.02754 (дата обращения: 24.10.2026).
5. Scikit-learn. RandomForestRegressor. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (дата обращения: 24.10.2026).
6. XGBoost. XGBoost Parameters. URL: https://xgboost.readthedocs.io/en/stable/parameter.html (дата обращения: 24.10.2026).
7. XGBoost. Prediction. URL: https://xgboost.readthedocs.io/en/stable/prediction.html (дата обращения: 24.10.2026).