

DOI: <https://doi.org/10.37162/2618-9631-2026-3-82-98>

УДК 504.3.054+004.85:504.064

Идентификация с помощью моделей машинного обучения неблагоприятных для рассеивания загрязняющих веществ метеорологических условий

Е.В. Татаринovich, Д.В. Борисов, И.Н. Кузнецова

*Гидрометеорологический научно-исследовательский центр
Российской Федерации, г. Москва, Россия
e.tatarinovich@yandex.ru*

Впервые для идентификации неблагоприятных метеорологических условий (НМУ) применены методы машинного обучения. Для исследований выбран район Москвы – Останкино. В работе использованы данные измерений метеорологических характеристик на телебашне Останкино в слое до высоты 503 м, ретроспективные данные ERA5 о ветре на изобарической поверхности 925 гПа, сумме осадков за 3 часа и высоте пограничного слоя атмосферы, а также данные измерений концентраций загрязняющих веществ за период 2020–2025 гг. При построении моделей машинного обучения НМУ выявлены значимые предикторы, наиболее информативными среди них оказались скользящие средние за 3–6 ч разности температуры в слое 2–253 м, средняя за предшествующие 3–6 ч скорость ветра в слое 201–305 м. Разработана и протестирована локальная модель идентификации НМУ с использованием пяти алгоритмов машинного обучения. Установлено, что для задачи бинарной классификации НМУ ансамблевые методы (CatBoost, случайный лес) обеспечивают наилучший баланс между минимизацией пропусков НМУ и ложных тревог ($F1 = 0.62–0.64$). Значимые предикторы НМУ послужат основой для усовершенствования применяемых статистических методов прогнозирования НМУ.

Ключевые слова: НМУ, методы машинного обучения, локальная модель НМУ

Identification of adverse meteorological conditions for pollutant dispersion using machine learning models

E.V. Tatarinovich, D.V. Borisov, I.N. Kuznetsova

*Hydrometeorological Research Center of Russian Federation, Moscow, Russia
e.tatarinovich@yandex.ru*

Machine learning methods were used for the first time to identify adverse meteorological conditions (AMC). The Ostankino district of Moscow was selected for the study. The study utilized meteorological measurement data from the Ostankino TV tower in a layer up to a height of 503 m, retrospective ERA5 data on wind at the 925 hPa level, 3-hour precipitation totals, and the height of the atmospheric boundary layer, as well as pollutant concentration measurements for the period of 2020–2025. When constructing the AMC machine learning models, significant predictors were identified, the most informative of which were 3–6-hour moving averages of the temperature difference in the 2–253 m layer and the average wind speed in the 201–305 m layer over the previous

3–6 hours. A local AMC identification model was developed and tested using five machine learning algorithms. It was found for the binary classification of atmospheric conditions that ensemble methods (CatBoost, random forest) provide the best balance between minimizing missed AMC and minimizing false alarms ($F1 = 0.62\text{--}0.64$). Significant predictors of AMC will serve as a basis for improving the statistical methods used for AMC forecasting.

Ключевые слова: adverse meteorological conditions, machine learning methods, local AMC model

Введение

Прогнозирование неблагоприятных для рассеивания загрязняющих веществ метеорологических условий (НМУ) в соответствии с Федеральным Законом № 96 «Об охране атмосферного воздуха» выполняется в оперативных подразделениях Росгидромета [9]. Гидрометцентр России в рамках действующей технологии осуществляет метеорологическое сопровождение работ по прогнозированию НМУ общего назначения. В комплект предоставляемой продукции входят прогностические значения метеорологического показателя рассеивания загрязнения (МППЗ), его расчет основан на учете типа температурной стратификации и скорости ветра в нижнем 250-метровом слое атмосферы, количества осадков в каждый час текущих и следующих суток [5, 6, 8]. При разработке МППЗ (более 7 лет назад) выбор названных выше трех предикторов был связан не только с ограниченным набором и объемом доступных данных метеорологических наблюдений в атмосферном пограничном слое, но и с учетом возможности использования их прогностических значений по расчетам численных моделей атмосферы для всей территории Российской Федерации.

Быстрое развитие численных моделей, увеличение пространственного и временного разрешения [7] позволяют применить более широкий набор метеорологических предикторов для прогнозирования МППЗ, в т. ч. в диапазоне МППЗ, соответствующем НМУ. Новые перспективы комплексного учета влияния значимых факторов открывает обращение к современным методам анализа с использованием машинного обучения, позволяющим использовать нелинейные взаимосвязи между факторами, извлекая закономерности непосредственно из данных без ручных настроек, в отличие от традиционных регрессионных методов [10]. В ряде научных публикаций, например [15], отмечается, что современные нейросетевые подходы обладают существенными преимуществами перед традиционными подходами в точности, гибкости и устойчивости моделирования природных процессов. В [21] описано применение машинного обучения для моделирования динамики городского теплового острова, применение глубокого обучения в области гидрологии иллюстрируется в [19]. Методы машинного обучения активно используются в области прогнозирования воздушного загрязнения [14, 16, 20, 22], эффективность их применения при прогнозировании концентраций PM_{10} и приземного озона в Москве показана в публикациях авторов [1, 2]. В [11] демонстрируется улучшение прогноза качества

воздуха в Европе, предоставляемого службой мониторинга атмосферы Copernicus, благодаря использованию методов машинного обучения. Даже небольшое число приведенных выше ссылок на публикации, посвященные методам машинного обучения, указывает на целесообразность их применения для повышения качества идентификации и прогноза метеорологических условий, способствующих загрязнению приземного воздуха, входящих в область профессиональных интересов авторов.

Целями данного исследования являются: а) первичная апробация методов машинного обучения; б) уточнение и/или установление решающих метеорологических факторов, обуславливающих повышение уровня загрязнения в мегаполисе. Такая постановка задачи ставится исходя из концепции связи роста уровня загрязнения в условиях слабого турбулентного и динамического перемешивания при отсутствии осадков, принятых обозначать как неблагоприятные метеорологические условия (НМУ). При решении задачи конкретизируется диапазон НМУ по индикаторам загрязнения приземного воздуха, а именно, по концентрациям наиболее чувствительных к метеоусловиям двух загрязняющих веществ – угарного газа (СО) и оксида азота (NO).

Конечной целью данного исследования является создание модели машинного обучения (ММО) идентификации НМУ (иНМУ), при которой выполняется условие одновременного превышения заданного порога концентрациями СО и NO, а именно, превышения 90-го перцентиля распределения концентраций, определенного по полной выборке за период 2020–2025 гг. С учётом сезонных вариаций измеренных концентраций значение 90-го перцентиля СО составило 600 мкг/м^3 (400 мкг/м^3 в декабре–марте), NO – 45 мкг/м^3 (30 мкг/м^3 в апреле–июле). Определённое таким образом число случаев иНМУ за пять лет составило 7 %. Заметим, что при ориентации на 90-й перцентиль использованы рекомендации ВМО [13], где предлагается явления с частотой более 90-го перцентиля определять как экстремальные, в нашей интерпретации – иНМУ.

В данной работе мы ограничились представлением разработки локальной ММО для района Останкино в Москве. Выбор этой локации определился наличием уникальных данных метеорологических наблюдений на телебашне Останкино (далее – ОТБ) до высоты 503 м; кроме того, рядом с ОТБ проводятся измерения концентраций загрязняющих веществ на автоматизированной станции контроля загрязнения атмосферы (АСКЗА) ГПБУ «Мосэкомониторинг» (<https://mosecom.mos.ru/air-quality/>).

Разработанная в ходе исследований модель машинного обучения представляет собой статистическую модель, обучаемую устанавливать значение бинарной целевой переменной иНМУ (1 – есть НМУ, 0 – нет НМУ) из набора числовых предикторов. В терминологии машинного обучения событием является состояние атмосферы, способствующее накоплению загрязняющих веществ с достижением и превышением критических уровней.

Использованные данные

В качестве признакового описания ММО использованы часовые данные измерений температуры, скорости ветра на ОТБ и рассчитанные по ним параметры, в т. ч. апробированный авторами ранее метеорологический параметр MP . MP рассчитывается за каждый час как произведение средней скорости ветра по измерениям на высотах 201, 253, 305 м ОТБ и разности температур в слое 2–253 м, нормированной на высоту 253 м. Для учёта предыдущего состояния атмосферы, влияющего на интенсивность рассеивания примесей, использованы агрегированные признаки и лаги – значения со сдвигом по времени назад. Идея работы с временными приращениями уже показала свою эффективность в других задачах, например в моделировании острова тепла [21]. Полный список и обозначения всех предикторов представлен в табл. 1.

Таблица 1. Предикторы, их обозначение и целевая переменная ММО, сгруппированные по виду и источникам сбора данных

Table 1. Machine learning model predictors and the target variable, grouped by data type and collection source

Источник данных	Тип данных	Предикторы
Метеонаблюдения на ОТБ и рассчитанные по ним параметры	Числовой	Температура воздуха на уровне 2 м, T_2 (°C) Разности температур воздуха в слое 2–253 м, dT_{253} ; в слое 2–503 м, dT_{503} (°C) Средняя скорость ветра на уровнях 201, 253, 305 м, U_{mean} (м/с) Параметр MP (°C м/с)
Реанализ ERA5	Числовой	Сумма осадков за 3 часа, $precip_3h$ (м) Высота погранслоя, h_{pbl} (м) Скорость ветра на и. п. 925 гПа, V_{925} (м/с) Направление ветра на и. п. 925 гПа, $wdir_{925_sin}$, $wdir_{925_cos}$ (радиан)
Календарные предикторы	Числовой	День года, $year_sin$, $year_cos$ (радиан) Час суток, day_sin , day_cos (радиан)
Агрегированные признаки	Числовой	Скользящие средние за 3, 6, 12 и 24 ч для dT_{253} (°C), параметра MP (°C м/с) и U_{mean} (м/с) Лаги для dT_{253} , Δ (°C) за 3, 6, 12 и 24 ч
Целевая переменная (по наблюдениям CO и NO)	Категориальный	иНМУ («есть/нет»)

Дополнительно к данным ОТБ в исследовательскую базу данных включены данные реанализа ERA5 в ближайшем к ОТБ узле сетки (на расстоянии 5,4 км): об осадках за час (в исследовании суммированные за 3 часа), высоте атмосферного пограничного слоя, скорости и направлении

ветра на изобарической поверхности (и. п.) 925 гПа [12]. База данных содержит выраженные в радианах календарные предикторы: день года, час суток, а также данные измерений концентрации CO и NO на АСКЗА Останкино.

Исследовательская база данных содержит информацию за период 1 августа 2020 г. – 31 июля 2025 г., без пропусков количество часовых записей составило 41806.

Методы исследования

Для решения поставленной задачи бинарной классификации НМУ на основе подготовленной базы данных было построено пять моделей машинного обучения различной архитектуры:

- логистическая регрессия;
- дерево решений;
- ансамбль деревьев решений *случайный лес*;
- алгоритм градиентного бустинга на деревьях решений (реализация CatBoost);
- искусственная нейронная сеть *многослойный перцептрон*.

Изучение возможностей и особенностей такого количества ММО определено желанием найти оптимальную ММО для практического использования. Подробному описанию ММО посвящен целый ряд публикаций последних лет [3, 17, 18], ниже отражены главные характеристики использованных ММО.

Логистическая регрессия использована в качестве базовой модели для оценки линейной классификации НМУ, в которой вероятность принадлежности объекта к классу рассчитывается с помощью логистической функции (сигмоиды), а параметры модели оцениваются методом максимального правдоподобия, что эквивалентно минимизации функции потерь (log-loss) [17]:

$$J = -\frac{1}{N} \sum_{i=1}^N [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)], \quad (1)$$

где N – количество объектов в выборке; y_i – истинный класс объекта (принимает значения 0 или 1); p_i – предсказанная моделью вероятность того, что объект принадлежит к классу 1. Главное преимущество логистической регрессии, несмотря на ее простоту, состоит в интерпретации полученных связей признаков с целевой переменной. Ограничением такой модели является ее линейная природа и невозможность учесть взаимосвязи признаков друг с другом.

Дерево решений – нелинейная модель классификации, строящая иерархическую структуру правил путём рекурсивного разбиения пространства признаков. Оптимальное разбиение на каждом шаге находится по максимальному снижению энтропии (в данной работе – перекрестная энтропия)

$$J_{ce} = -\sum_{c \in Y} p_c^{(R)} \log p_c^{(R)}, \quad (2)$$

где $p_c^{(R)}$ — доля обучающих примеров класса c в листе R . Из недостатков «дерева решений» можно выделить склонность к переобучению, т. е. усвоение шума в обучающей выборке, что отрицательно сказывается на обобщающей способности модели на новых данных. К преимуществам можно отнести абсолютную интерпретируемость правил разбиения.

Случайный лес – ансамблевый метод, основанный на агрегировании в подходе бутстрэп множества деревьев решений [17]. Каждое дерево обучается на бутстрэп-выборке (порождённой одним и тем же распределением), а финальный прогноз формируется голосованием. Несмотря на то, что решение ансамбля сложнее интерпретировать по сравнению с деревом решений, этот подход борется с проблемой переобучения.

CatBoost (реализация градиентного бустинга) – метод ансамблевого обучения, который строит композицию из множества слабых моделей (деревьев решений), но в нем каждое следующее дерево исправляет ошибки предыдущих. Для предотвращения утечки информации и снижения переобучения градиенты функции потерь (в данном исследовании log-loss) для каждого объекта вычисляются только на основе предыдущих объектов во временном порядке. Для лучшей регуляризации, устойчивости к шуму и ускорения прогноза используются не обычные деревья, которые имеют разную структуру для разных ветвей, а симметричные, в которых все узлы на одном уровне используют один и тот же признак и порог. CatBoost считается одним из мировых стандартов для анализа табличных данных, позволяя эффективно обрабатывать нечисловые данные, не требуя их сложного предварительного кодирования. Благодаря встроенным методам борьбы со смещением данных, алгоритм устойчив к переобучению.

Многослойный перцептрон – искусственная нейронная сеть прямого распространения с одним или несколькими скрытыми слоями. В многослойном перцептроне вектор предикторов проходит через несколько линейных преобразований, обернутых нелинейной функцией активации (слои сети). Финальный слой преобразует вектор в скаляр (прогноз сети). Обучение модели выполняется методом градиентного спуска, на каждом шаге вычисляются градиенты функции потерь (binary cross-entropy для задачи классификации) по параметрам линейных преобразований на слоях нейронной сети, параметры обновляются в сторону минимизации функции потерь.

Для моделей сложной архитектуры (*дерево решений, случайный лес, CatBoost, многослойный перцептрон*) дополнительно проводилась оптимизация гиперпараметров (табл. 2) по метрике качества F1 с использованием библиотеки Optuna. Гиперпараметры – это внешние настройки модели машинного обучения, которые задаются разработчиком до начала процесса обучения и не изменяются самой моделью в отличие от внутренних параметров (например, весов в нейросетях), которые алгоритм вычисляет сам в процессе обучения.

Таблица 2. Оптимизированные гиперпараметры моделей
Table 2. Optimized hyperparameters of machine learning models

Модель	Оптимизированные гиперпараметры
Дерево решений	Максимальная глубина дерева: 13 Минимальное количество наблюдений для разбиения узла: 34 Минимальное количество наблюдений в листе: 31 Критерий разбиения: энтропия Параметр пост-прунинга: 0.00117 Минимальное уменьшение неопределённости: 3.14e-05 Вес классов: включен
Случайный лес	Количество деревьев: 422 Максимальная глубина дерева: 21 Минимальное количество наблюдений для разбиения узла: 6 Минимальное количество наблюдений в листе: 8 Максимальное число признаков для разбиения: квадратный корень из общего числа признаков Критерий разбиения: энтропия Минимальное уменьшение неопределённости: 2.19×10^{-6} Вес классов: включен
Catboost	Максимальная глубина дерева: 9 Количество деревьев: 703 Скорость обучения: 0.0605 L2-регуляризация листьев: 14.21 Сила случайности: 8.98 Температура бэггинга: 0.374 Число границ (интервалов): 74 Вес положительного класса: включен
Многослойный перцептрон	Размеры скрытых слоев: 96 и 16 нейронов Функция активации: ReLU Оптимизатор: SGD Инерция направления обновления весов: 0.93 L2-регуляризация: $\alpha=0.0131$ Размер батча: 128 Стратегия скорости обучения: адаптивная, начиная с 0.00217 Число эпох: 300 с ранней остановкой Доля валидации: 15 %

Стоит отметить, что в *логистической регрессии*, *дереве решений*, *случайном лесе* и *Catboost* дисбаланс целевой переменной (в нашем случае 7 % случаев *иНМУ* в наборе данных) учитывался с помощью гиперпараметров моделей, а именно, весов классов. Заметим, что в *многослойном перцептрон*е протестированные стратегии обработки дисбаланса классов (*RandomUnderSampler* и *SMOTE*) не привели к улучшению качества идентификации *иНМУ*, поэтому балансировку классов оптимальная конфигурация не включала.

Сравнение качества разработанных ММО проводилось традиционным способом – с использованием таблицы сопряженности и метрик качества, включающих попадания, пропуски и ложные тревоги иНМУ (табл. 3).

Таблица 3. Матрица сопряжённости

Table 3. Contingency matrix

Факт	0	Оправдавшееся отсутствие явления <i>TN</i> (True Negative)	Ложные тревоги <i>FP</i> (False Positive)
	1	Пропуски <i>FN</i> (False Negative)	Попадания <i>TP</i> (True Positive)
		0	1
		Прогноз	

Так как иНМУ – явление редкое, рассматривались следующие метрики:

– предупредённость $Recall = \frac{TP}{TP+FN}$ – чтобы учесть пропуски;

• $Precision = \frac{TP}{TP+FP}$ – чтобы учесть ложные тревоги;

• $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$ – гармоническое среднее между ними;

• $F2$ – аналогично $F1$, только с большим весом $Recall$:

$$F2 = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}, \beta = 2, \text{ так как пропустить}$$

НМУ критичнее, чем спрогнозировать ложную тревогу;

• Критерий надёжности Багрова – Хайдке для оценки качества по сравнению со случайным прогнозом:

$$HSS = \frac{2(TP \cdot TN) - (FP \cdot FN)}{[(TP + FN) \cdot (FN + TN) + (TP + FP) \cdot (FP + TN)]}$$

Поскольку метрика $F1$ обеспечивает баланс между минимизацией пропусков и ложных тревог, она была выбрана основной для оптимизации гиперпараметров моделей и оценки перестановочной важности признаков.

Для обеспечения сопоставимости результатов ММО различной архитектуры оценка важности отдельных предикторов выполнялась с использованием перестановочного подхода для всех моделей, включая логистическую регрессию. Данный метод устанавливает изменение целевой метрики ($F1$) при случайном перемешивании значений признака на тестовой выборке (в данной работе – на 5 тестовых фолдах кросс-валидации, т. е. на подмножествах данных), что позволяет оценить важность признака для прогноза целевой переменной. Для оценки неопределённости метрик качества и важности признаков в подходе кросс-валидации применялось годовое разбиение данных, в котором в качестве фолда выступал один метеорологический год, т. е. для используемого пятилетнего набора данных формировалось 5 выборок, в каждой из которых обучение осуществлялось на 4 годах (обучающая выборка) и проверка – на одном году (тестовая

выборка). Такой подход представляется оптимальным для поставленной задачи, так как позволяет учесть не только автокорреляцию, характерную для метеорологических данных, но и суточный ход и сезонные особенности, поскольку соблюдается баланс сезонных и суточных циклов в обучающей и тестовой выборках, а также гарантируется, что в тестовую выборку не попадут значения, сильно коррелированные со значениями в обучающей выборке (соседние часы).

Алгоритмы формирования обучающей и тестовой выборок, конструирования, обучения и тестирования моделей машинного обучения были разработаны на языке программирования Python с использованием библиотек для работы с данными Numpy, Pandas и библиотеки машинного обучения scikit-learn.

Обсуждение результатов

По целевой метрике F1, под максимизацию которой оптимизировались все модели сложной архитектуры, наилучший результат составил у ММО CatBoost $F1=0.638 \pm 0.017$, при этом превосходство CatBoost над случайным лесом и многослойным перцептроном не достигло уровня статистической значимости (табл. 4).

Таблица 4. Сравнительные метрики качества моделей на тестовой выборке ($\pm$ стандартное отклонение), разность F1 на обучающей и тестовой выборках ($\Delta F1$)

Table 4. Comparative model quality metrics on the test set ($\pm$ standard deviation), F1 difference between the training and test sets ($\Delta F1$)

Метрики качества моделей	Модель				
	Логистическая регрессия	Дерево решений	Случайный лес	CatBoost	Многослойный перцептрон
Precision	0.360 ± 0.036	0.337 ± 0.038	0.562 ± 0.041	0.574 ± 0.047	0.673 ± 0.043
Recall	0.885 ± 0.037	0.868 ± 0.017	0.719 ± 0.037	0.724 ± 0.039	0.543 ± 0.089
F1	0.510 ± 0.030	0.484 ± 0.038	0.629 ± 0.013	0.638 ± 0.017	0.595 ± 0.044
F2	0.682 ± 0.011	0.657 ± 0.026	0.679 ± 0.016	0.686 ± 0.017	0.561 ± 0.070
HSS	0.4455 ± 0.0333	0.4148 ± 0.0417	0.5917 ± 0.0125	0.6016 ± 0.0175	0.5635 ± 0.0474
$\Delta F1$	0.017	0.063	0.217	0.284	0.033

В табл. 4 видно, что разработанные ММО демонстрируют различные стратегии балансировки между Precision и Recall, что отражает их архитектурные особенности и целевые функции оптимизации:

– Логистическая регрессия и дерево решений показывают высокий Recall (> 0.86) при низком Precision (≈ 0.34 – 0.36). Это означает, что модели минимизируют пропуски НМУ, но прогнозируют значительное количество ложных тревог. Такая стратегия приводит к невысоким значениям F1 (0.48 – 0.51).

– Ансамблевые модели (случайный лес, CatBoost) достигают наилучших значений показателя F1 (0.62–0.64) за счёт более равномерного распределения важности между признаками (рис. 1) и способности улавливать нелинейные зависимости. При этом наблюдается компромисс: более высокая точность Precision (≈ 0.56 –0.57) сопровождается умеренным Recall (≈ 0.72). Вместе с тем в ансамблевых моделях, несмотря на применение техник регуляризации, наиболее заметен разрыв качества на обучающих данных и тестовой выборке ($\Delta F1 = 0.21$ –0.28), что указывает на переобучение моделей на специфических паттернах обучающих выборок (паттерны в этом случае – закономерности в данных, которые модель выявляет в процессе обучения и использует для классификации).

– Многослойный перцептрон демонстрирует наиболее сбалансированные значения Precision (0.673) и Recall (0.543), что обеспечивает относительно высокий F1 (0.595) при минимальном переобучении ($\Delta F1 = 0.033$). Однако высокая вариабельность результатов ($\sigma = 0.044$) ограничивает надёжность этой ММО.

Исходя из полученных результатов, можно сделать вывод, что лучшим вариантом из разработанных ММО для идентификации иНМУ является CatBoost, сочетая наивысшее значение F1 (0.638) с высокой устойчивостью ($\sigma = 0.017$). Случайный лес демонстрирует сопоставимое качество (F1 = 0.629) при несколько меньшей устойчивости и умеренном переобучении, поэтому может рассматриваться как надёжная альтернатива при ограниченных вычислительных ресурсах. Многослойный перцептрон, также показавший удовлетворительный результат (F1 = 0.595), но с высокой нестабильностью, в текущей конфигурации требует дополнительной настройки. Аналогичные выводы следуют при оценке качества разработанных ММО с использованием критерия надёжности Багрова – Хайдке (HSS): наибольший HSS=0.6 имеет ММО CatBoost.

Как результат первого подхода к решению задачи идентификации НМУ на основе методов машинного обучения приведенные выше показатели качества разработанных ММО следует признать удовлетворительными, учитывая и редкость целевых событий, и достаточно строгое принятое условие иНМУ – одновременное превышение критических порогов концентрации CO и NO. При этом следует отметить, что при условном постоянстве антропогенных выбросов в мегаполисе [4] в колебаниях уровня городского загрязнения присутствуют не связанные с метеорологическими условиями скачки, вызванные особенностями режима городского транспорта (основного источника загрязнения в городе). В их числе – увеличение транспортных потоков в тёплый период в так называемый «дачный сезон» с пятницы до понедельника, загрязнение воздуха при массовом выезде снегоуборочной техники в периоды сильных снегопадов, связанные с авариями или строительными работами длительные заторы на крупных автомагистралях, внеплановые выбросы крупных промышленных предприятий и др.

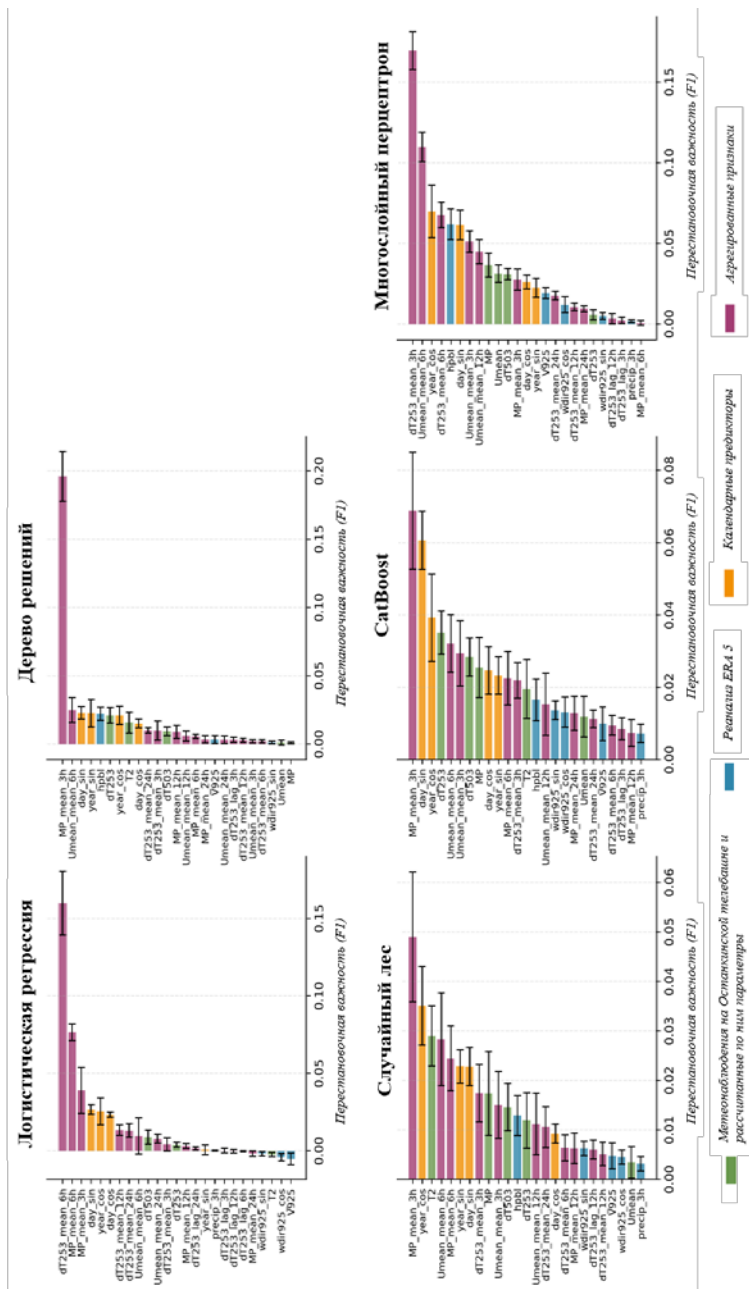


Рис. 1. Перестановочная важность признаков (по метрике F1), полосами погрешностей обозначены 95% доверительные интервалы. Для наглядности для каждой ММО отображены 25 верхних признаков. Цветом обозначены группы признаков в соответствии с табл. 1.

Fig. 1. Permutation feature importance (based on the F1 metric); error bars indicate 95% confidence intervals. For clarity, the top 25 features are displayed for each ML model. Feature groups are color-coded in accordance with Table 1

Подобные нерегулярные всплески концентраций загрязняющих веществ, безусловно, создают дополнительные шумы, препятствуя максимально достижимому качеству моделей, но пока устойчивая статистика их проявления отсутствует, и авторы не ставили задачу получить ее вследствие ограниченной выборки данных (5 лет) и имеющейся тенденции снижения уровня городского загрязнения в последние годы [4].

Сопутствующий разработке ММО анализ значимости признаков позволил на основе перестановочной их важности выявить ранее не использованные в схеме прогноза НМУ предикторы. В табл. 5 приводятся наиболее информативные для каждой разработанной ММО предикторы – первые 5 в ранжированном ряду предикторов, которые содержат ключевые метеорологические характеристики:

- разность температуры в слое 2–253 м и скользящие средние за 3 и 6 ч (dT_{253} , $dT_{253_mean_3h}$, $dT_{253_mean_6h}$);
- скользящие средние за предшествующие 3 и 6 ч значения параметра MP (MP_mean_3h , MP_mean_6h);
- скользящее среднее за 6 ч осредненной по нескольким высотам скорости ветра на Останкинской телебашне ($U_{mean_mean_6h}$);
- календарные предикторы (день года, час суток).

Как видно, все модели улавливают важность предыдущего состояния атмосферы, отдавая предпочтение скользящим средним значениям за 3–6 ч диапазон, только в двух моделях в число топ-5 признаков входят не смещенные по времени метеорологические параметры ($hpbl$ в ММО дерево решений и $T2$ в ММО случайный лес). Календарные признаки, отражая имеющие место суточные и сезонные циклы метеорологической обусловленности процессов повышения уровня загрязнения, вносят небольшой, но стабильный вклад во всех моделях, за исключением ММО CatBoost, в которой признак day_sin лишь немного уступает приоритетному признаку MP_mean_3h .

На рис. 1 приводятся по 25 «верхних» признаков перестановочной важности по метрике $F1$ для каждой ММО, которые, учитывая корреляцию между ними, можно воспринимать как методическую основу для выбора предикторов при усовершенствовании статистических методов идентификации и прогнозирования НМУ при изменении целевой переменной и планируемого перехода от локальной модели к региональной.

Заключение

Впервые для идентификации неблагоприятных метеорологических условий (НМУ) применены методы машинного обучения. Для исследования выбран район на севере Москвы, где, кроме наземных измерений метеорологических характеристик и концентраций загрязняющих веществ, проводятся метеорологические наблюдения на 8 уровнях в слое до 500 м на телебашне Останкино.

Таблица 5. Топ-5 признаков каждой ММО (в скобках – значение перестановочной важности, т. е. разность значений целевой метрики F1 до и после случайного перемешивания значений признака)

Table 5. Top 5 predictors for each machine learning model (in parentheses: permutation importance value – i. e., the difference in the target F1 metric before and after randomly shuffling the feature values)

Модель	Топ-5 признаков	Интерпретация
Логистическая регрессия	dT253_mean_6h (0.160) MP_mean_6h (0.077) MP_mean_3h (0.039) day_sin (0.027) year_cos (0.026)	Линейная модель наиболее чувствительна к скользящим средним разности температуры и параметра MP
Дерево решений	MP_mean_3h (0.196) Umean_mean_6h (0.025) day_sin (0.023) year_sin (0.023) hpbl (0.022)	Дерево выделяет скользящее среднее параметра MP за 3 ч как доминирующий предиктор, возможно, отражая пороговый характер связи MP с НМУ, а также – склонность одиночного дерева к созданию жёстких правил на основе одного главного признака
Случайный лес	MP_mean_3h (0.049) year_cos (0.035) T2 (0.029) Umean_mean_6h (0.028) MP_mean_6h (0.024)	Ансамбль деревьев распределяет важность более равномерно, что согласуется с её способностью усреднять множественные слабые сигналы, кроме MP выделяя сезонность (year_cos), температуру на 2 м и ветер как дополнительные предикторы
CatBoost	MP_mean_3h (0.069) day_sin (0.061) year_cos (0.039) dT253 (0.035) Umean_mean_6h (0.032)	Градиентный бустинг кроме параметра MP акцентирует внимание на суточных (day_sin) и годовых (year_cos) циклах, что может указывать на нелинейные временные паттерны
Многослойный перцептрон	dT253_mean_3h (0.170) Umean_mean_6h (0.110) year_cos (0.070) dT253_mean_6h (0.068) hpbl (0.062)	Нейронная сеть отдаёт преимущество разности температуры в слое и ветру, возможно, улавливая сложные нелинейные взаимодействия между признаками

На основе сформированной пятилетней выборки (август 2020 г. – июль 2025 г.) протестированы пять моделей машинного обучения (ММО) различной архитектуры. Установлено, что для задачи бинарной классификации НМУ ансамблевые методы (CatBoost, случайный лес) обеспечивают наилучший баланс между минимизацией пропусков и ложных тревог ($F1 = 0.62\text{--}0.64$). Качество исследуемых ММО в виду отсутствия аналогов для сравнений следует признать удовлетворительным.

При анализе обученных моделей выявлены наиболее информативные предикторы сформулированных при постановке задачи НМУ; в их число вошли скользящие средние за 3–6 ч разности температуры в слое 2–253 м и метеорологического параметра *MP*, а также осредненной по высотам ОТБ скорости ветра, которые отражают основные физические механизмы переноса и накопления загрязняющих веществ в приземном воздухе.

Результаты проведенного исследования могут стать методической основой для дальнейшей разработки методов идентификации и прогнозирования НМУ на основе машинного обучения. В перспективе планируется использовать полученные знания о наиболее важных признаках формирования неблагоприятных для рассеивания метеорологических условий при переходе от локальной модели к региональной и подготовке для неё базы данных.

Авторы выражают благодарность сотрудникам ГПБУ «Мосэкоммониторинг» Департамента природопользования г. Москвы за предоставление данных и многолетнее сотрудничество в области мониторинга окружающей среды.

Отдельно выражаем благодарность М.И. Варенцову и М.А. Криницкому за введение в науку машинного обучения и ценные консультации.

Статья подготовлена с использованием УНУ «ИТК ГМП», код: 4071013.

Список литературы

1. Борисов Д.В., Кузнецова И.Н. Постпроцессинг численных прогнозов концентраций приземного озона с использованием машинного обучения // Гидрометеорологические исследования и прогнозы. 2023. № 4 (390). С. 86–104. DOI: <https://doi.org/10.37162/2618-9631-2023-4-86-104>
2. Борисов Д.В., Кузнецова И.Н. Интегрирование химической транспортной модели и искусственной нейронной сети для прогноза концентрации PM10 // Гидрометеорологические исследования и прогнозы. 2024. № 3 (393). С. 42–63. DOI: <https://doi.org/10.37162/2618-9631-2024-3-42-63>
3. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science : Пер. с англ. СПб.: БХВ-Петербург, 2022. 352 с.
4. Департамент природопользования и охраны окружающей среды города Москвы Доклад о состоянии окружающей среды в городе Москве в 2024 году [Электронный ресурс]. URL: <https://www.mos.ru/eco/function/departament/doklady-o-deyatelnosti/>.
5. Кузнецова И.Н., Ткачева Ю.В., Шалыгина И.Ю., Нахаев М.И. Прогнозирование метеорологического показателя рассеивания загрязняющих веществ в приземном воздухе // Гидрометеорологические исследования и прогнозы. 2021. № 3 (381). С. 131–149. DOI: <https://doi.org/10.37162/2618-9631-2021-3-131-149>
6. Кузнецова И.Н., Ткачёва Ю.В., Борисов Д.В. Методы прогнозирования метеорологических условий, влияющих на загрязнение приземного воздуха // Метеорология и гидрология. 2024. № 8. С. 87–103. DOI: 10.52002/0130-2906-2024-8-87-103
7. Ривин Г.С., Розинкина И.А., Астахова Е.Д., Блинов Д.В. и др. Система краткосрочного численного прогноза высокой детализации COSMO-Ru, ее развитие и приложения // Гидрометеорологические исследования и прогнозы. 2019. № 374 (4). С. 37–53.

8. Ткачева Ю.В., Кузнецова И.Н., Нахаев М.И. Усовершенствованная технология прогноза метеорологических параметров и условий, влияющих на загрязнение воздуха // Гидрометеорологические исследования и прогнозы. 2022. № 3 (385). С. 161–177. DOI: <https://doi.org/10.37162/2618-9631-2022-3-161-177>
9. Федеральный закон от 04.05.1999 № 96-ФЗ. Об охране атмосферного воздуха [Электронный ресурс]: URL: https://www.consultant.ru/document/cons_doc_LAW_22971.
10. Шолле Ф. Глубокое обучение на Python. СПб.: Питер, 2018. 400 с.
11. Bertrand J.-M., Meleux F., Ung A., Descombes G., Colette A. Technical note: Improving the European air quality forecast of the Copernicus Atmosphere Monitoring Service using machine learning techniques // *Atmospheric Chemistry and Physics*. 2023. Vol. 23, no. 9. P. 5317–5333. <https://doi.org/10.5194/acp-23-5317-2023>
12. Copernicus Climate Change Service (C3S) at ECMWF ERA5: Fifth generation ECMWF atmospheric reanalysis of the global climate [Электронный ресурс]. URL: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels>.
13. Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation // WCDMP-No. 72. Geneva: World Meteorological Organization, 2009. 55 с.
14. Kayastha S.G., Ghahremanloo M., Park J., Singh D., Westenbarger D., Choi Y. A Deep Learning Framework for Satellite-Derived Surface PM_{2.5} Estimation: Enhancing Spatial Analysis in the United States // *Artificial Intelligence for the Earth Systems*. 2024. Vol. 3. DOI: 10.1175/AIES-D-24-0028.1
15. Krinitskiy M.A. An Overview of Machine Learning and Deep Learning Applications in Earth Sciences in 2024: Achievements and Perspectives // *Moscow University Physics Bulletin*. 2024. Vol. 79, no. S2. P. S739–S749. DOI: 10.3103/S0027134924702217
16. Masood A., Ahmad K. A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance // *Journal of Cleaner Production*. 2021. Vol. 322, no. 11. P. 129072. DOI: 10.1016/j.jclepro.2021.129072
17. Murphy K.P. Probabilistic machine learning: an introduction. Cambridge, Massachusetts, The MIT Press, 2022. 826 с.
18. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features // *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. 2018. C. 6638–6649.
19. Shen C., Lawson K. Applications of deep learning in hydrology // *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*. 2021. C. 283–297.
20. Suslov A.I., Krinitskiy M.A., Staquet C., Le Boudec E. Machine Learning Methods for Statistical Prediction of PM_{2.5} in Urban Agglomerations with Complex Terrain, Using Grenoble As an Example // *Moscow Univ. Phys. Bulletin*. 2024. Vol. 79, no. S2. P. S774–S783. DOI: 10.3103/S0027134924702242
21. Varentsov M., Krinitskiy M., Stepanenko V. Machine Learning for Simulation of Urban Heat Island Dynamics Based on Large-Scale Meteorological Conditions // *Climate*. 2023. Vol. 11, no. 10. P. 200. <https://doi.org/10.3390/cli11100200>
22. Wang S., McGibbon J., Zhang Y. Predicting high-resolution air quality using machine learning: Integration of large eddy simulation and urban morphology data // *Environmental Pollution*. 2024. Vol. 344. <https://doi.org/10.1016/j.envpol.2024.123371>.

References

1. Borisov D.V., Kuznetsova I.N. Postprocessing chislennyh prognozov koncentracij prizemnogo ozona s ispol'zovaniem mashinnogo obucheniya [Post-processing of ground-level ozone numerical forecasts using machine learning]. *Gidrometeorologicheskie issledovaniya i prognozy* [*Hydrometeorological Research and Forecasting*], 2023, no. 4 (390), pp. 86–104. <https://doi.org/10.37162/2618-9631-2023-4-86-104> [in Russ.].
2. Borisov D.V., Kuznetsova I.N. Integrirovaniye himicheskoy transportnoy modeli i iskusstvennoy nejronnoy seti dlya prognoza koncentracii PM₁₀ [Integration of a chemical transport

model and an artificial neural network for PM₁₀ concentration forecasting]. *Gidrometeorologicheskie issledovaniya i prognozy [Hydrometeorological Research and Forecasting]*, 2024, no. 3 (393), pp. 42-63. <https://doi.org/10.37162/2618-9631-2024-3-42-63> [in Russ.].

3. Bryus P., Bryus E. *Prakticheskaya statistika dlya specialistov Data Science [Practical statistics for Data Scientists]*. Sankt-Peterburg, BHV-Peterburg publ., 2022, 352 p. [in Russ.].

4. Departament prirodopol'zovaniya i ohrany okruzhayushchej sredy goroda Moskvy Doklad o sostoyanii okruzhayushchej sredy v gorode Moskve v 2024 godu [Elektronnyj resurs]. URL: <https://www.mos.ru/eco/function/departament/doklady-o-deyatelnosti/> [in Russ.].

5. Kuznetsova I.N., Tkacheva Yu.V., Shalygina I.Yu., Nakhaev M.I. Prognozirovanie meteorologicheskogo pokazatelya rasseivaniya zagryaznyayushchih veshchestv v prizemnom vozduhe. [Forecasting a meteorological indicator of pollutant dispersion in surface air]. *Gidrometeorologicheskie issledovaniya i prognozy [Hydrometeorological Research and Forecasting]*, 2021, no. 3 (381), pp. 131-149. DOI: <https://doi.org/10.37162/2618-9631-2021-3-131-149> [in Russ.].

6. Kuznetsova I.N., Tkacheva Yu.V., Borisov D.V. Methods for forecasting meteorological conditions affecting surface air pollution. *Russ. Meteorol. Hydrol.*, 2024, vol. 49, no. 8, pp. 722-734. DOI: [10.3103/S1068373924080077](https://doi.org/10.3103/S1068373924080077)

7. Rivin G.S., Rozinkina I.A., Astakhova E.D., Blinov D.V. et al. Sistema kratkosrochnogo chislennogo prognoza vysokoj detalizacii COSMO-Ru, ee razvitie i prilozheniya [COSMO-Ru high-resolution short-range numerical weather prediction system: its development and applications]. *Gidrometeorologicheskie issledovaniya i prognozy [Hydrometeorological Research and Forecasting]*, 2019, no. 4 (374), pp. 37-53. [in Russ.].

8. Tkacheva Yu.V., Kuznetsova I.N., Nakhaev M.I. Usovershenstvovannaya tekhnologiya prognoza meteorologicheskikh parametrov i uslovij, vliyayushchih na zagryaznenie vozduha [Improved technology for forecasting meteorological parameters and conditions affecting air pollution]. *Gidrometeorologicheskie issledovaniya i prognozy [Hydrometeorological Research and Forecasting]*, 2022, no. 3 (385), pp. 161-177. DOI: <https://doi.org/10.37162/2618-9631-2022-3-161-177> [in Russ.].

9. Federal'nyj zakon ot 04.05.1999 № 96-ФЗ. Ob ohrane atmosfernogo vozduha [Elektronnyj resurs] : URL: https://www.consultant.ru/document/cons_doc_LAW_22971.

10. Sholle F. *Deep Learning with Python*. Sankt-Peterburg, Piter publ., 2018, 400 p. [in Russ.].

11. Bertrand J.-M., Meleux F., Ung A., Descombes G., Colette A. Technical note: Improving the European air quality forecast of the Copernicus Atmosphere Monitoring Service using machine learning techniques. *Atmospheric Chemistry and Physics*, 2023, vol. 23, no. 9, pp. 5317-5333. <https://doi.org/10.5194/acp-23-5317-2023>

12. Copernicus Climate Change Service (C3S) at ECMWF ERA5: Fifth generation ECMWF atmospheric reanalysis of the global climate [Электронный ресурс]. URL: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels>.

13. Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation. WCDMP-No. 72, Geneva, World Meteorological Organization, 2009, 55 p.

14. Kayastha S.G., Ghahremanloo M., Park J., Singh D., Westenbarger D., Choi Y. A Deep Learning Framework for Satellite-Derived Surface PM_{2.5} Estimation: Enhancing Spatial Analysis in the United States. *Artificial Intelligence for the Earth Systems*. 2024. Vol. 3. DOI: [10.1175/AIES-D-24-0028.1](https://doi.org/10.1175/AIES-D-24-0028.1)

15. Krinitskiy M.A. An Overview of Machine Learning and Deep Learning Applications in Earth Sciences in 2024: Achievements and Perspectives. *Moscow University Physics Bulletin*. 2024, vol. 79, no. S2, pp. S739-S749. DOI: [10.3103/S0027134924702217](https://doi.org/10.3103/S0027134924702217)

16. Masood A., Ahmad K. A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of Cleaner Production*, 2021, vol. 322, no. 11, pp. 129072. DOI: [10.1016/j.jclepro.2021.129072](https://doi.org/10.1016/j.jclepro.2021.129072)

17. Murphy K.P. *Probabilistic machine learning: an introduction*. Cambridge, Massachusetts, The MIT Press publ., 2022. 826 c.

18. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features // *Advances in Neural Information Processing Systems* 31 (NeurIPS 2018), 2018, pp. 6638-6649.

19. Shen C., Lawson K. Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 2021, pp. 283-297.

20. Suslov A.I., Krinitskiy M.A., Staquet C., Le Boudec E. Machine Learning Methods for Statistical Prediction of PM2.5 in Urban Agglomerations with Complex Terrain, Using Grenoble As an Example. *Moscow Univ. Phys. Bulletin*, 2024, vol. 79, no. S2, pp. S774-S783. DOI: 10.3103/S0027134924702242

21. Varentsov M., Krinitskiy M., Stepanenko V. Machine Learning for Simulation of Urban Heat Island Dynamics Based on Large-Scale Meteorological Conditions. *Climate*, 2023, vol. 11, no. 10, pp. 200. <https://doi.org/10.3390/cli11100200>

22. Wang S., McGibbon J., Zhang Y. Predicting high-resolution air quality using machine learning: Integration of large eddy simulation and urban morphology data // *Environmental Pollution*. 2024. Vol. 344. <https://doi.org/10.1016/j.envpol.2024.123371>.

*Поступила 06.05.2026; принята в печать 22.10.2026.
Submitted 06.05.2026; accepted for publication 22.10.2026.*