СРАВНЕНИЕ АНСАМБЛЕВЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ПРИ РЕШЕНИИ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ ОКОНЧАНИЯ ПЕРИОДА ЗАМОРОЗКОВ

В. А. Солозобов[1,а], **С. А. Лысенкова**[1,б]
[1] *Сургутский государственный университет, г. Сургут, Российская Федерация*
[а] ✍ *solo.val.al@yandex.ru*
[б] *lsa1108@mail.ru*

Аннотация: в статье приводится сравнение результатов применения ансамблевых методов машинного обучения для решения задачи прогнозирования завершения периода заморозков. Дано краткое описание ансамблевых методов. Представлены результаты исследования зависимости влияния различных наборов гиперпараметров и входных данных на обучение оптимальной модели. Сделаны выводы о качестве получаемых моделей с помощью различных вариаций градиентного бустинга, случайного леса и линейной модели. В работе приведены результаты применения библиотек, реализующих методы машинного обучения: XGBoost, LightGBM, CatBoost, Random Forest (scikit-learn), логистическая регрессия.

Ключевые слова: ансамблевые методы машинного обучения, градиентный бустинг, период заморозков, прогноз заморозков, временные ряды.

Для цитирования: Солозобов В. А., Лысенкова С. А. Сравнение ансамблевых методов машинного обучения при решении задачи прогнозирования окончания периода заморозков. *Успехи кибернетики.* 2026;7(2):132–138.

Поступила в редакцию: 05.03.2026. *В окончательном варианте: 22.03.2026.*

COMPARISON OF ENSEMBLE MACHINE LEARNING METHODS FOR PREDICTING THE END OF THE FROST PERIOD

V. A. Solozobov[1,a], **S. A. Lysenkova**[1,b]
[1] *Surgut State University, Surgut, Russian Federation*
[а] ✍ *solo.val.al@yandex.ru*
[б] *lsa1108@mail.ru*

Abstract: we studied the performance of ensemble machine learning methods for predicting the end of the frost period. We provided a brief overview of the considered ensemble approaches and investigated how different hyperparameter settings and input data configurations affect model training. We applied several tools, including gradient boosting methods (XGBoost, LightGBM, and CatBoost), random forest (scikit-learn), and logistic regression (scikit-learn), compared the resulting models, and assessed their predictive quality. The results show differences in performance across methods and highlight the impact of hyperparameter tuning and input data selection on prediction accuracy.

Keywords: ensemble machine learning methods, gradient boosting, frost period, frost prediction, time series.

Cite this article: Solozobov V. A., Lysenkova S. A. Comparison of Ensemble Machine Learning Methods for Predicting the End of the Frost Period. *Russian Journal of Cybernetics.* 2026;7(2):132–138.

Original article submitted: 05.03.2026. *Revision submitted: 22.03.2026.*

Постановка решаемой задачи

Решаемая в данной работе задача была описана ранее [1]. В кратком изложении она заключается в прогнозировании дня, после которого наступает устойчивая положительная температура в весенне-летний период. Сам прогноз планируется осуществлять каждый день, и тот день, когда модель выдаст высокую прогнозную вероятность окончания периода заморозков, будет означать, что наступил период стабильно положительной температуры и понижение ее ниже нуля не ожидается вплоть до осени. При этом модель будет решать задачу бинарной классификации временных рядов метеопараметров, где метка класса «1» относится к случаю, когда заморозки еще возможны, а «0» — к случаю, когда заморозков не ожидается вплоть до осени. Расчет прогноза модель

будет осуществлять на временных рядах различных метеопараметров. Ранее данная задача была решена с помощью логистической регрессии, в данном случае планируется использовать ансамблевые методы машинного обучения (МО) для того, чтобы увеличить точность прогнозирования. Также остаются вопросы выбора оптимального набора входных данных, то есть того, какие метеопараметры и какой длины временное окно требуются, чтобы получить наилучшую точность при обучении модели.

Целью исследования является решение задачи ансамблевыми методами машинного обучения и сравнение эффективности и качества получаемых моделей между собой, а также с теми, что получаются при использовании методов логистической регрессии.

В рамках исследования рассматривается использование библиотек, реализующих ансамблевые методы МО: XGBoost, LightGBM, CatBoost, Random Forest (scikit-learn). А также анализируется влияние признаков и их количества на метрики качества получаемых моделей.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Изучить особенности ансамблевых методов машинного обучения. Рассмотреть основные, широко используемые библиотеки, реализующие построение ансамблей, проанализировать отличительные характеристики, достоинства и недостатки.
2. Определить критерии качества, по которым будет происходить сравнение полученных моделей, подготовить варианты группировки входных данных, провести обучение моделей.
3. Провести сравнение полученных моделей, определить закономерности и влияние набора входных данных, выделить особенности. Сделать общий вывод по результатам исследования и определить дальнейшее направление по улучшению качества модели прогнозирования окончания периода заморозков.

Модель будет обучаться на метеорологических данных Сургутского района [2]. Данная задача весьма актуальна для данного региона из-за нестабильного весеннего периода и довольно сильного разброса дат последнего морозного дня весной от года к году. Для обучения модели были выбраны следующие метеопараметры [3]:

температура воздуха (градусы Цельсия) на высоте 2 метра над поверхностью земли (T);

атмосферное давление на уровне станции (миллиметры ртутного столба) (P);

относительная влажность (%) на высоте 2 метра над поверхностью земли (H);

направление ветра (румбы) на высоте 10–12 метров над земной поверхностью, осредненное за 10-минутный период (Wd);

скорость ветра на высоте 10–12 метров над земной поверхностью, осредненная за 10-минутный период (метры в секунду) (Ws);

общая облачность (%) (Cp);

температура точки росы на высоте 2 метра над поверхностью земли (градусы Цельсия) (Dw).

Данные метеопараметры будут включены в обучающие выборки как отдельно, так и в различных комбинациях. Также обучающие выборки будут различаться по длине временного окна от 1 до 120 суток.

Одной из главных проблем при решении поставленной задачи является выбор признаков. Он определяется как метеопараметрами, которые включаются в обучающую выборку, так и длиной временного ряда. При этом увеличение количества признаков не всегда ведет к повышению качества прогноза. В зависимости от применяемых алгоритмов МО, часто выбор большого числа признаков приводит к нестабильным результатам прогнозирования, даже при учете переобучения моделей. Также происходит увеличение времени на обучение модели.

В данном исследовании поставлена задача изучить влияние длины временных рядов (временного окна) и участвующих метеопараметров на результативность обучения следующего ряда ансамблевых методов МО: XGBoost, LightGBM, CatBoost, Random Forest. Также в сравнении с ними будут участвовать модели, построенные методом логистической регрессии. Результативность обучения моделей будет оцениваться метрикой качества, определенной в работе [1]. Эта метрика показывает среднее количество общих ошибок модели, приходящихся на один год. Отбор моделей и гиперпараметров, таких как глубина деревьев, коэффициент скорости обучения модели и другие, будет проводиться по валидационной выборке. При этом будет происходить калибровка

порогового значения, разделяющего объекты на классы, таким образом, чтобы ложноотрицательная ошибка на обучающей и валидационной выборках была равна нулю. Это делается по причине критичности ошибки неверного прогноза наступления теплой погоды: ее следует привести к нулю. И уже по тестовой выборке будет происходить окончательная оценка откалиброванных моделей.

Описание ансамблевых методов машинного обучения

Одним из направлений развития методов машинного обучения (МО) можно назвать ансамблирование моделей, полученных методами МО. Ансамбль моделей подразумевает наличие некоторого множества базовых моделей (БМ), каждая из которых выдает свое решение по поводу поставленной общей задачи, после чего на основе полученного множества решений высчитывается конечный результат работы данного ансамбля. Важно отметить, что для получения ансамбля с высокими предсказательными качествами требуется сильное разнообразие базовых моделей [4]. Под разнообразием понимается различие признаков и зависимостей, которые описывают базовые модели.

Методы построения базовых моделей можно классифицировать по зависимости построения базовых моделей друг от друга [5]. Зависимое обучение базовых моделей обусловлено тем, что для получения новой БМ требуется получить результат обучения предыдущей БМ, то есть ансамбль приходится строить последовательно. При независимом обучении БМ возможно параллельное построение ансамбля, что позволяет ускорить данный процесс.

Основным методом, реализующим параллельный подход, является алгоритм случайного леса (Random Forest) [6]. В этом алгоритме при построении базовых деревьев, из которых состоит случайный лес, вводится случайный выбор набора признаков, по которым идет разделение в узле решающего дерева. За счет этого удается сильно увеличить разнообразие получаемых БМ. Следует выделить следующие преимущества алгоритма случайного леса: благодаря методу обучения сборки, случайные леса могут уменьшить дисперсию и повысить стабильность прогнозов по сравнению с моделями индивидуальной классификации, использование метода построения деревьев на разных и независимых подмножествах снижает проблему переобучения и улучшает способность к обобщению [7]. Алгоритм случайного леса хорошо работает с различными большими и сложными данными и демонстрирует хорошую способность справляться с пропущенными данными и помехами. Предоставляет оценку значимости признаков, позволяя пользователям определить наиболее важные признаки модели и понять относительное влияние каждого признака на результаты [8]. И все это с высокой скоростью обучения ансамбля и возможностью распараллеливания процесса.

Ярким примером последовательного обучения является подход градиентного бустинга, основанный на том, что каждая последующая обучаемая БМ исправляет ошибку всех БМ до нее. За счет этого разрешается также и проблема расчета итогового результата [9, 10]. Одной из первых успешных реализаций градиентного бустинга в машинном обучении можно назвать алгоритмы библиотеки XGBoost. Она представляет собой масштабируемые алгоритмы, используемые для задач классификации и регрессии с применением регуляризации, которая предотвращает переобучение [11]. Стоит отметить следующие преимущества XGBoost: способность обрабатывать пропущенные значения, возможность минимальной обработки признаков, таких как нормализация данных и масштабирование признаков.

Следующей известной библиотекой, реализующей алгоритмы градиентного бустинга, можно назвать LightGBM [12]. Она использует в качестве базовых моделей деревья решений и применяется, в основном, для задач классификации, ранжирования. Библиотека LightGBM отличается повышенной скоростью обучения и меньшими требованиями к памяти при обучении по сравнению с другими реализациями градиентного бустинга. Это преимущество достигается за счет ряда алгоритмических решений в области подготовки данных к обучению. Происходит исключение данных, которые не вносят большой вклад в градиент ошибки, часть данных сжимается, где это возможно без потери информации и где признаки носят зачастую нулевые значения [13]. Также свою роль играют алгоритмы бинаризации непрерывных величин у признаков. В свою очередь, эти манипуляции с обучающим набором могут приводить к переобучению на небольших наборах данных [12].

Еще одна успешная библиотека, реализующая алгоритмы градиентного бустинга, — это

CatBoost. Одним из главных нововведений CatBoost является его способность выполнять несмещенную оценку градиента, что снижает переобучение, а также упрощает конечную ансамблевую модель и ускоряет расчеты обученного ансамбля. Достигается это за счет симметрии выстраиваемых деревьев в базовых моделях. Заметным преимуществом библиотеки CatBoost являются алгоритмы автоматического преобразования категориальных признаков [14].

Можно отметить следующие общие свойства, характерные для библиотек, реализующих алгоритмы градиентного бустинга. По результату их применения есть возможность определить значимость признаков, что может быть использовано для их отбора. Каждая рассмотренная реализация градиентного бустинга позволяет избежать переобучения [9]: за счет скорости обучения, возможности обучаться на небольших выборках и с неоднородными данными. Градиентный бустинг получил широкое применение при работе с табличными данными [15, 16]. Есть положительные результаты использования данного метода при обучении на многомерных временных рядах [16].В результате проведенного обзора решено использовать ансамбли градиентного бустинга (XGBoost, LightGBM, CatBoost) и Random Forest как представителя параллельного ансамбля. Данные реализации выбраны как имеющие хорошие показатели точности получаемых моделей и широкие возможности настройки обучения. Предлагается сравнить их работу при решении поставленной ранее задачи как между собой, так и с классическим линейным классификатором.

Результаты

В первую очередь, были получены метрики качества на моделях, построенных на данных, которые включали временной ряд только одного из метеопараметров (рис. 1). По ним можно сделать вывод о том, что, используя такие параметры, как температура или точка росы, достаточно размера временного окна в 10 дней. При этом модели имеют значения среднегодовой ошибки на уровне 10 и потом она уже незначительно снижается по мере увеличения временного окна. Это можно объяснить тем, что период заморозков и его окончание имеют непосредственное отношение к температуре воздуха и по тому, какой температурный режим был в последние 10–30 дней, уже можно судить о вероятности окончания периода заморозков. При этом увеличение временного окна для температуры до 2–3 месяцев уже не приносит ощутимого увеличения точности в прогнозирование, а даже мешает прогнозированию, внося свою долю шума, на которую приходится обращать внимание обучающим алгоритмам. В отличие от температуры и точки росы, обучение на других метеопараметрах требует временного окна не менее 60 дней и наименьший уровень ошибок достигается при дальнейшем его увеличении (рис. 1). То есть 1–2 месяцев данных недостаточно для получения точной модели по этим метеопараметрам. Третьим метеопараметром после температуры и точки росы по влиянию на точность получаемых моделей является относительная влажность, но, как было отмечено, для этого требуется временное окно не менее 60 дней (рис. 1). На одиночных параметрах хорошо себя показывают библиотеки, реализующие градиентный бустинг. XGBoost и LightGBM показывают схожесть по точности получаемых моделей. CatBoost при этом обучает модели, схожие по качеству как с XGBoost, так и с Random Forest.

У Random Forest, в свою очередь, получаются более хорошие результаты, чем у других методов, на метеопараметрах: температура, сила ветра, направление ветра. Логистическая регрессия показывает себя хуже, особенно тяжело по сравнению с другими дается прогноз по давлению и проценту облачности.

Следующим этапом было сравнение моделей, обученных на полном наборе метеопараметров, с обученными на одном временном ряду температуры и добавление к нему других метеопараметров поочередно (рис. 2).

Линейная модель довольно ярко иллюстрирует, как добавление каждого нового параметра приводит к уменьшению ошибки, но это происходит только при временных окнах до 40 дней. При этом наилучшие показатели у небольшого набора из 4 метеопараметров. После отметки в 40 дней происходит рост средней ошибки, при этом наибольшее ухудшение у моделей с большим числом метеопараметров. Подобное наблюдается и у ансамбля Random Forest. Что касается ансамблей градиентного бустинга, то влияние количества метеопараметров незначительно сказывается на ошибке модели. Но важно отметить, что модели, обученные на среднем наборе метеопараметров (3–4), все же выше по точности, чем построенные на одном метеопараметре или на всех сразу. При

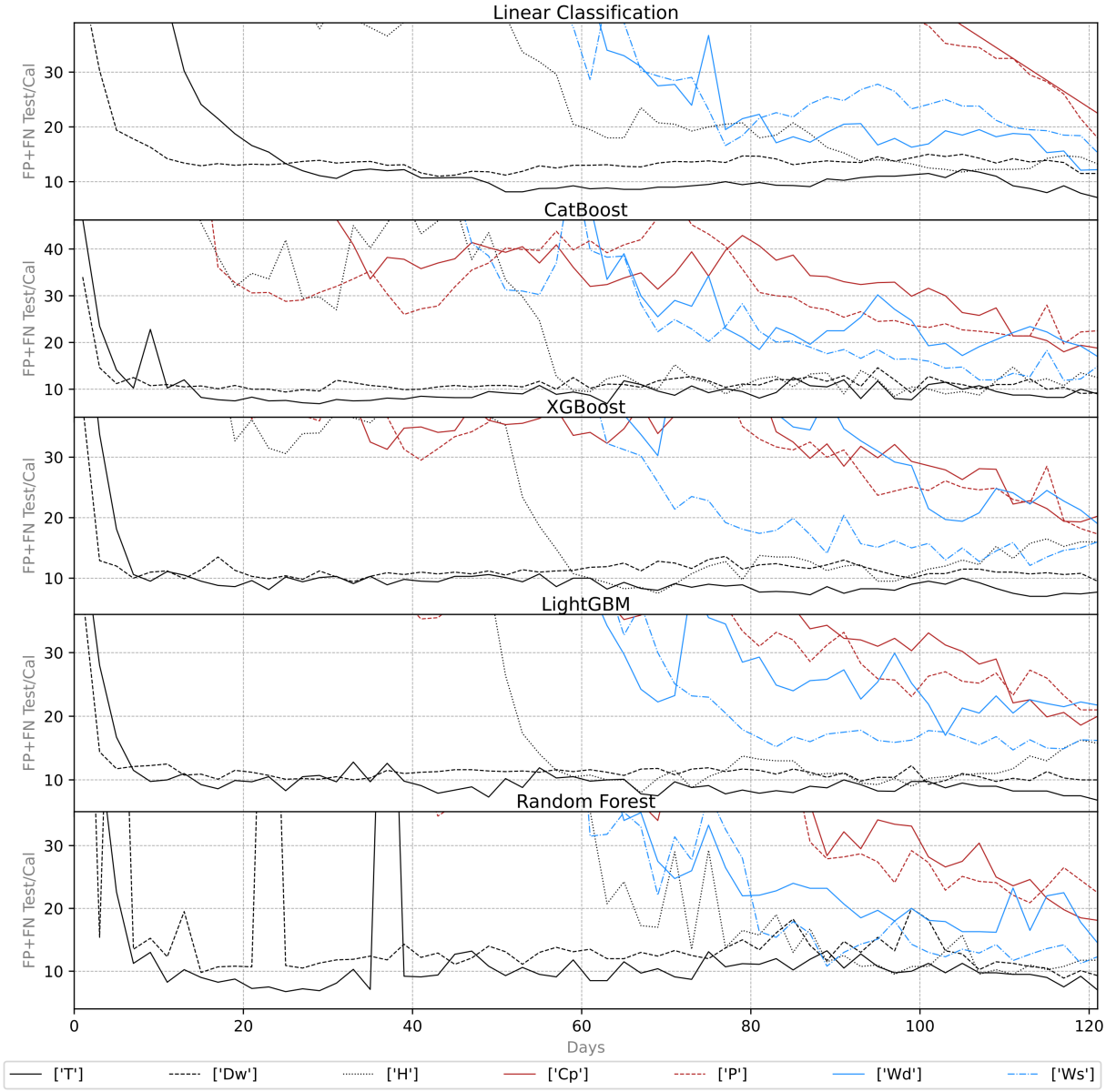

Рис. 1. *Сравнение графиков влияния одиночных признаков и величины временного окна на метрику качества моделей, получаемых различными МО. FP+FN Test/Cal — общая среднегодовая ошибка по тестовой выборке после калибровки порогового значения. T — температура воздуха; P — давление воздуха; Dw — точка росы; Cp — процент облачности; Ws — скорость ветра; Wd — направление и скорость ветра; H — относительная влажность воздуха*

этом линейная модель в итоге получается с самой низкой ошибкой (FN + FP = 4,5), за ней идет по результативности алгоритм Random Forest. То есть с помощью логистической регрессии удается получить наилучший результат. Но для этого результата требуются определенные условия: 30-40 дней временное окно и набор метеопараметров (T, Wd, Cp, H). При других конфигурациях обучающей выборки логистическая регрессия дает результаты хуже ансамблевых. Из ансамблевых методов можно отметить два: XGBoost и CatBoost. Они дают стабильные результаты и, в среднем, более низкую ошибку, чем другие.

Заключение

В результате работы была получена модель с прогнозной среднегодовой ошибкой, равной 4,5. Но получена она методами логистической регрессии. Наилучшие результаты по методам МО получились следующие: Random Forest — 5,25, CatBoost — 5,7, XGBoost — 5,9, LightGBM — 6,4.

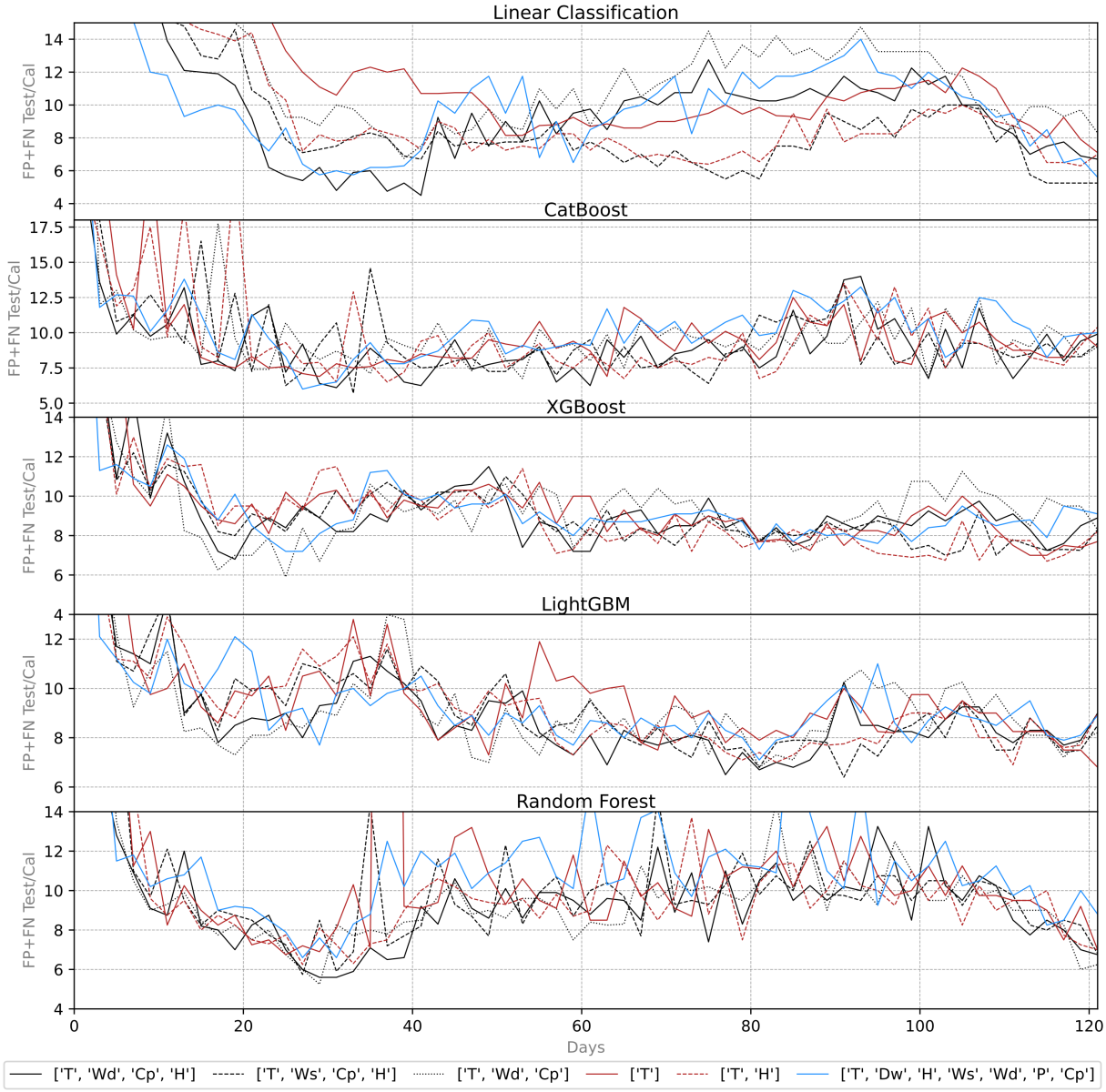

Рис. 2. *Сравнение графиков влияния различных наборов признаков и величины временного окна на метрику качества моделей, получаемых различными МО. FP+FN Test/Cal — общая среднегодовая ошибка по тестовой выборке после калибровки порогового значения. T — температура воздуха; P — давление воздуха; Dw — точка росы; Cp — процент облачности; Ws — скорость ветра; Wd — направление и скорость ветра; H — относительная влажность воздуха*

Это не значит, что ансамблевые методы хуже, они дают более стабильный результат на большем диапазоне возможных вариантов подготовки обучающей выборки, чем логистическая регрессия. Стоит также отметить закономерность для всех методов МО, используемых в данной работе: наиболее точные модели получаются при обучающей выборке, состоящей из следующих метеопараметров: T, Wd, Cp, H, и в диапазоне временного окна от 20 до 40 суток, также наблюдается качественное уменьшение ошибок моделей, полученных при временном окне в диапазоне 110–120 суток.

По данному результату можно сделать следующие выводы. Добавление новых метеопараметров в обучающую выборку приводит к серьезному увеличению объема входных данных, что приводит к некоторому роду зашумленности данных, и зависимости, которые выявлялись ранее, перестают определяться: они теряются на фоне других вероятных признаков. При этом добавление нового временного ряда, в котором явно есть признаки, по которым можно прогнозировать

окончание периода заморозков, не приносит уменьшения ошибки. Поэтому следует решить проблему избыточности данных. Для признаков температуры и точки росы можно, начиная с длины временного окна в 30 дней, уже использовать не все замеры, а усредненные по суткам, возможно, с учетом минимальных и максимальных значений. Другие метеопараметры, вероятно, совсем не требуют такого количества замеров, и усреднение может производиться за 5–10 дней. Также серьезное сокращение входных данных можно получить за счет приобретения новых признаков на основе имеющихся, заменяя их таким образом.

В районе значений временного окна, равных 120 дням, заметна тенденция к уменьшению ошибки. Возможно, следует в обучающую выборку добавить информацию за период прошлой осени – начала зимы, в ней могут быть признаки окончания периода заморозков.

Еще одним выходом из ситуации большого числа признаков может быть метод построения ансамбля из ансамблей. При этом каждый базовый ансамбль обучается на своем одном метеопараметре и со своим временным окном. Далее результаты работы этих ансамблей будут агрегироваться в простейшем варианте логистической регрессией.

ЛИТЕРАТУРА

1. Солозобов В. А., Лысенкова С. А. Методика оценки и выбора оптимальной модели, прогнозирующей окончание периода заморозков. *Успехи кибернетики.* 2025;6(2):100–107. EDN: WKIVRC.
2. *Архив метеоданных.* Режим доступа: https://rp5.ru/.
3. Рубан А. А., Белогубова А. А. Прогнозирование выручки предприятий общественного питания с использованием прогноза погоды. *Финансовый вестник.* 2025;1:20–25. EDN: VFGNRG.
4. Rokach L. Ensemble-Based Classifiers. *Artificial Intelligence Review.* 2010;33(1):1–39. DOI: 10.1007/s10462-009-9124-7.
5. Yang Y., Lv H., Chen N. A Survey on Ensemble Learning under the Era of Deep Learning. *Artificial Intelligence Review.* 2023;56:5545–5589. DOI: 10.1007/s10462-022-10283-5.
6. Schonlau M., Zou R. Y. The Random Forest Algorithm for Statistical Learning. *The Stata Journal.* 2020;20(1):3–29. DOI: 10.1177/1536867X20909688.
7. Biau G., Scornet E. A Random Forest Guided Tour. *Test.* 2016;25(2):197–227. DOI: 10.1007/s11749-016-0481-7.
8. Salman H. A., Kalakech A., Steiti A. Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning.* 2024;2024:69–79. DOI: 10.58496/BJML/2024/007.
9. Natekin A., Knoll A. Gradient Boosting Machines, a Tutorial. *Frontiers in Neurorobotics.* 2013;7:21. DOI: 10.3389/fnbot.2013.00021.
10. Bentéjac C., Csörgő A., Martínez-Muñoz G. A Comparative Analysis of Gradient Boosting Algorithms. *Artificial Intelligence Review.* 2021;54(3):1937–1967. DOI: 10.1007/s10462-020-09896-5.
11. Chen T., Guestrin C. Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIKGDD International Conference on Knowledge Discovery and Data Mining.* 2016:785–794. DOI: 10.1145/2939672.2939785.
12. Ke G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems.* 2017:30.
13. Zhang Y., Liu J., Shen W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences.* 2022;12(17):8654. DOI: 10.3390/app12178654.
14. Prokhorenkova L. et al. CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems.* 2018:31.
15. Hancock J. T., Khoshgoftaar T. M. CatBoost for Big Data: an Interdisciplinary Review. *J Big Data.* 2020;7:94. DOI: 10.1186/s40537-020-00369-8.
16. Jiang J. et al. Boosting Tree-Assisted Multitask Deep Learning for Small Scientific Satasets. *Journal of Chemical Information and Modeling.* 2020;60(3):1235–1244. DOI: 10.1021/acs.jcim.9b01184.