УДК 004.932.2:004.896

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АНСАМБЛЕВЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАБОЛЕВАЕМОСТИ ГРИППОМ

Насибуллин М.А.[1]

Магистрант

Казанский национальный исследовательский технический университет им. А.Н. Туполева - КАИ

Россия, Казань

Аннотация. В статье рассматриваются ансамблевые методы машинного обучения на основе деревьев решений. Используя алгоритмы Random Forest (случайный лес) и CatBoost (градиентный бустинг), выполняется прогнозирование заболеваемости гриппом. В исследовании используются данные эпидемиологического мониторинга, предоставляемые Всемирной организацией здравоохранения (ВОЗ) в рамках глобальной системы эпиднадзора за гриппом FLuNet. В качестве входных данных используются признаки, построенные на основе эпидемиологических наблюдений, включая лаговые значения заболеваемости. Качество моделей оценивается с использованием различных метрик. Результаты позволяют сделать оценку эффективности рассматриваемых алгоритмов и определить более подходящий подход для прогнозирования эпидемиологических показателей.

Ключевые слова: ансамблевые методы, машинное обучение, прогнозирование временных рядов, грипп, эпидемиологические данные.

COMPARATIVE ANALYSIS OF ENSEMBLE MACHINE LEARNING METHODS FOR INFLUENZA INCIDENCE FORECASTING

Nasibullin M.A.

[1] *Научный руководитель – к.т.н., доцент Кремлева Э.Ш.*

Graduate Student
Kazan National Research Technical University named after A.N. Tupolev-KAI
Russia, Kazan

Abstract. This article explores ensemble machine learning methods based on decision trees. Using the Random Forest and CatBoost algorithms, influenza incidence forecasting is performed. The study utilizes epidemiological surveillance data provided by the World Health Organization (WHO) within the Global Influenza Surveillance and Response System (GISRS), specifically the FluNet platform. The input features are constructed based on epidemiological observations, including lagged incidence values. Model performance is evaluated using various metrics. The results allow assessing the effectiveness of the considered algorithms and identifying a more suitable approach for forecasting epidemiological indicators.
Keywords: ensemble methods, machine learning, time series forecasting, influenza, epidemiological data.

Введение

Грипп остаётся одной из наиболее распространённых острых респираторных вирусных инфекций, оказывающих существенное влияние на здоровье населения, что регулярно формирует существенную нагрузку на системы здравоохранения. В периоды эпидемического подъёма наблюдается увеличение числа обращений за медицинской помощью, рост госпитализаций, а также косвенные экономические потери, связанные с временной утратой трудоспособности населения. В связи с этим задача прогнозирования заболеваемости гриппом является актуальным направлением исследований в области эпидемиологии и анализа данных.

Развитие глобальных систем эпидемиологического мониторинга позволило сформировать крупные массивы данных о распространении

гриппа. Одним из ключевых источников таких данных является система FluNet, входящая в состав Global Influenza Surveillance and Response System (GISRS) Всемирной организации здравоохранения, которая обеспечивает сбор и публикацию данных о лабораторно подтверждённых случаях гриппа в различных странах мира [1].

В последние годы для анализа эпидемиологических временных рядов всё шире применяются методы машинного обучения, позволяющие учитывать нелинейные зависимости и сложную структуру данных. Современные учебные и методические материалы по машинному обучению подтверждают эффективность применения алгоритмов ансамблевого обучения для задач прогнозирования и анализа данных [2, 3].

Особый интерес представляют ансамблевые методы, такие как Random Forest и градиентный бустинг, которые показали высокую эффективность в задачах регрессии и прогнозирования на табличных данных. Теоретические основы этих методов подробно рассмотрены в классических работах и учебных курсах по машинному обучению [4, 5].

Целью данной работы является сравнительный анализ ансамблевых алгоритмов машинного обучения [8] для прогнозирования заболеваемости гриппом на основе данных эпидемиологического мониторинга FluNet. В рамках исследования рассматривается качество прогнозных моделей и проводится оценка их эффективности с использованием стандартных метрик регрессионного анализа.

Цель и задачи исследования

Целью данной работы является сравнительный анализ ансамблевых алгоритмов машинного обучения Random Forest и CatBoost для прогнозирования заболеваемости гриппом.

Для достижения цели решаются следующие задачи:

1. сбор и предварительная обработка данных
2. формирование лаговых признаков
3. формирование обучающей и тестовой выборки на основе временной структуры данных
4. обучение моделей Random Forest и CatBoost;
5. получение прогнозов
6. визуализация результатов
7. оценка качества с использованием метрик.

Данные исследования

Изначально имеются данные эпидемиологического мониторинга, предоставляемые ВОЗ. Рассматривается агрегированный временной ряд заболеваемости гриппом по России. Сходный набор состоит из следующих данных:

1. год наблюдения
2. номер недели
3. общее число зарегистрированных случаев гриппа

После выгрузки этих данных выполняется устранение отсутствующих значений: ячейки, где число зарегистрированных случаев гриппа отсутствует, заполняются нулями. Далее данные сортируются по временной шкале для формирования единого временного ряда.

Для прогнозирования этих данных недостаточно, поэтому добавляются следующие лаговые признаки, которые часто используются в задачах анализа временных рядов [7]:

1. lag_1 – число больных на прошлой неделе
2. lag_2 – число больных 2 недели назад
3. lag_3 – число больных 3 недели назад

4. lag_year – число больных на той же неделе предыдущего года

Таким образом, модель использует в качестве признаков номер недели, количество заболевших и представленные выше лаговые признаки. Модель будет обучаться на данных за 2016–2023 года и тестироваться на данных за 2024–2025 года.

Описание моделей и метрик

В данной работе были рассмотрены два ансамблевых метода машинного обучения: Random Forest и CatBoost. Оба этих метода относятся к семейству методов, основанных на решающих деревьях.

Random Forest основан на методе бэггинга. Модель строит множество решающих деревьев независимо друг от друга на случайном подмножестве признаков. Целевое значение вычисляется как среднее результатов всех деревьев. Основным преимуществом этого метода является устойчивость к переобучению и шуму в данных за счет снижения дисперсии.

CatBoost – метод градиентного бустинга над решающими деревьями [6]. В данном методе деревья строятся последовательно, обучаясь на ошибках предыдущих моделей, используя градиентный спуск. Ключевой особенностью является снижение смещения и предотвращение утечки информации.

Гиперпараметры для обоих моделей были подобраны эмпирически с учетом компромисса между точностью и устойчивостью моделей.

Для оценки качества моделей используются разные метрики регрессионного анализа.

MAE (Mean Absolute Error) – средняя абсолютная ошибка (1). Данная метрика показывает, насколько прогнозные значения в среднем отличаются от фактических. Для большей наглядности еще используется нормированная

версия MAE в процентах от диапазона значений.

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

RMSE (Root Mean Squared Error) – корень из средней квадратичной ошибки (2). Эта метрика тоже отражает среднюю величину ошибки, но она более чувствительна к большим отклонениям из-за квадратичного усиления.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

R^2 (коэффициент детерминации) – показатель качества аппроксимации, отражающий долю объяснённой дисперсии целевой переменной (3).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3)$$

Результаты

Программная реализация исследования выполнена на языке Python. Для обработки и анализа данных использовались библиотеки pandas и NumPy, для построения моделей машинного обучения – scikit-learn и CatBoost, а для визуализации результатов – matplotlib [2, 7]. Реализованная программа обеспечивает автоматическую загрузку эпидемиологических данных, формирование лаговых признаков, обучение моделей, построение прогнозов и расчёт показателей качества.

В результате выполнения программы были получены прогнозы заболеваемости гриппом на тестовой выборке за 2024–2025 годы. На рисунках 1 и 2 представлены графики с фактическими значениями и значениями, спрогнозированными моделями.

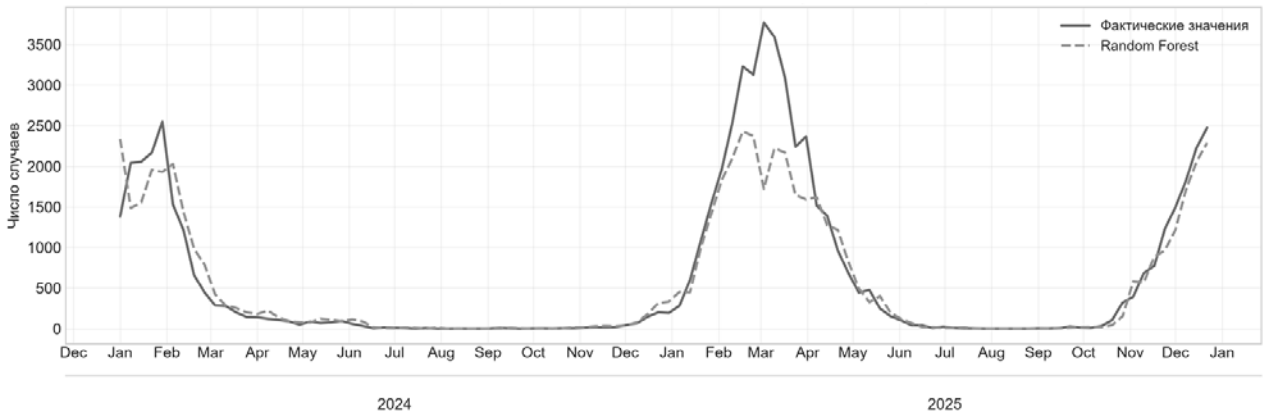

Рисунок 1. Сравнение фактических и прогнозных значений (Random Forest)

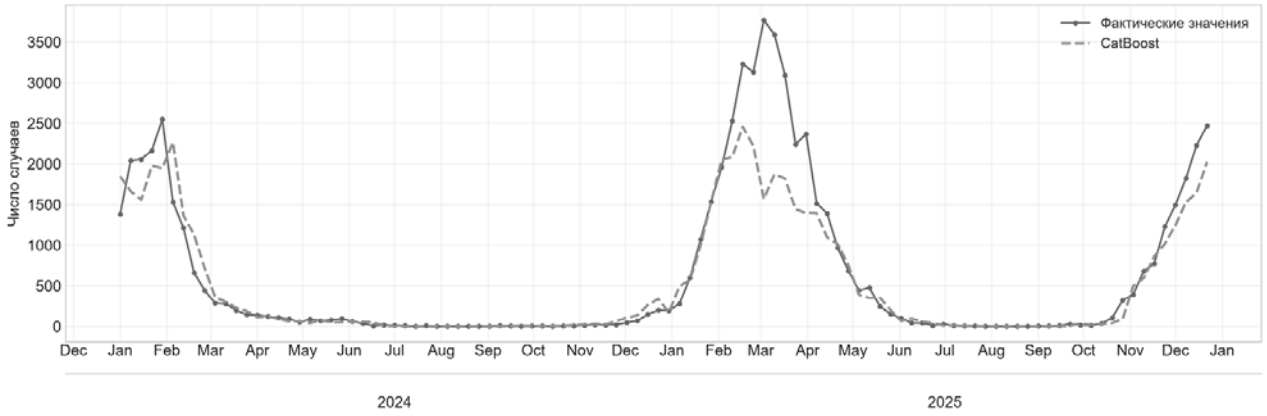

Рисунок 2. Сравнение фактических и прогнозных значений (CatBoost)

На рисунке 3 представлены диаграммы рассеяния фактических и прогнозных значений для обоих моделей.

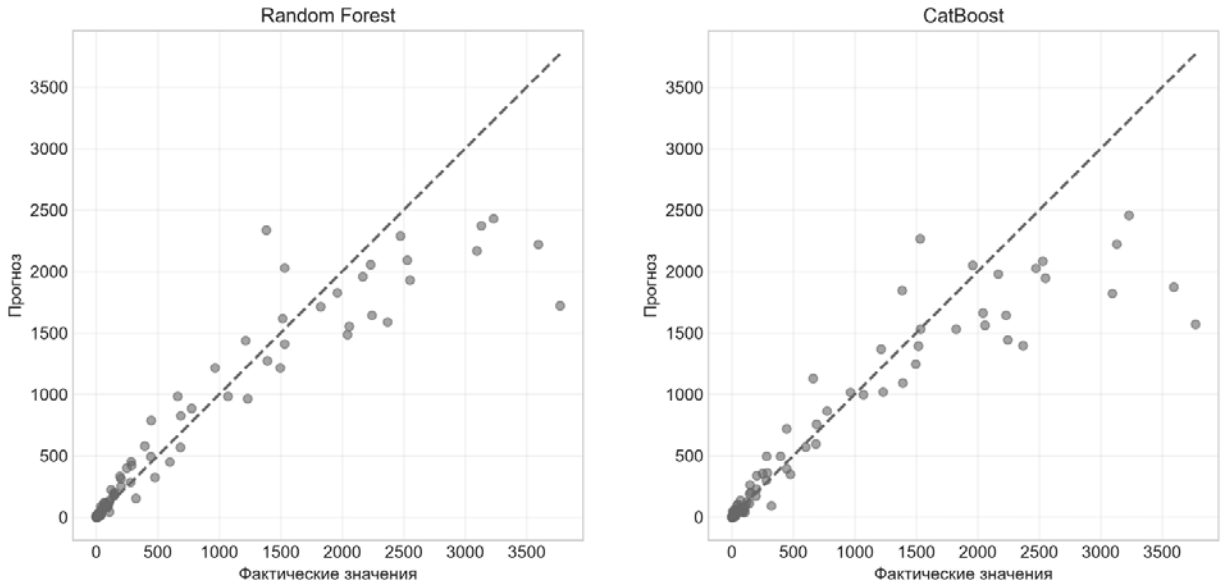

Рисунок 3. Диаграммы рассеяния.

Для оценки качества моделей рассчитывались метрики MAE, RMSE, MAE (% от диапазона значений) и коэффициент детерминации R^2. Полученные результаты представлены в таблице 1.

Таблица 1. Адаптация корректирующих действий под тип привода

Модель	MAE	RMSE	NMAE	R^2
RandomForest	158,7	344,7	4,21	0,87
CatBoost	170,2	386,73	4,52	0,84

В таблице $NMAE = \frac{MAE}{range} * 100\%$, где range – разница между максимальным и минимальным количеством заболевших в тестовом диапазоне. NMAE позволяет ошибку относительно масштаба временного ряда.

Исходя из метрик и графиков, можно сказать, что обе модели демонстрируют достаточно высокое качество прогнозирования заболеваемости гриппом. Тем не менее, в периоды пиковых значений наблюдается тенденция к недооценке экстремальных значений обеими моделями, что является типичной проблемой для моделей, обученных на

лаговых признаках без явного моделирования эпидемических всплесков.

При этом модель Random Forest показывает лучшие результаты по всем используемым в данной работе показателям качества: значение MAE составляет 158,7, RMSE — 344,7, коэффициент детерминации R^2 равен 0,87, а относительная ошибка составляет около 4,21% от диапазона значений. Это свидетельствует о высокой точности и устойчивости модели при прогнозировании временного ряда.

Заключение

В рамках данной работы проведен сравнительный анализ ансамблевых методов машинного обучения для задачи прогнозирования заболеваемости гриппом.

В процессе работы были выполнены этапы обработки данных и формирования необходимых признаков для обучающей модели, построения моделей Random Forest и CatBoost.

Судя по метрикам, можно сказать, что обе модели эффективно справились с задачей прогнозирования эпидемиологических данных, однако модель Random Forest показала более высокую точность.

Таким образом, можно сделать вывод, что ансамблевые методы машинного обучения являются хорошим инструментом для решения задач прогнозирования эпидемиологических временных рядов. Перспективным направлением дальнейших исследований является использование более сложных моделей глубокого обучения и учет дополнительных внешних факторов (погодные условия, сезонность, вакцинация).

Библиографический список

1. World Health Organization. Global Influenza Surveillance and Response System (GISRS) — [Электронный ресурс]. — Режим доступа — URL: https://www.who.int/initiatives/global-influenza-surveillance-and-response-system (Дата обращения 17.06.2026).
2. Школа анализа данных Яндекса. Онлайн-учебник по машинному обучению // Яндекс. — [Электронный ресурс]. — Режим доступа — URL: https://education.yandex.ru/handbook/ml (Дата обращения 17.06.2026)
3. Николенко С. И. Машинное обучение: основы. – М.: ДМК Пресс, 2025. – 608 с.
4. Шай Шалев-Шварц, Шай Бен-Давид. Идеи машинного обучения. От теории к алгоритмам. – М.: ДМК Пресс, 2019. – 436 с.
5. Мерков А. Б. Распознавание образов. Введение в методы статистического обучения. – М.: МГУ, 2011. – 256 с.
6. Дайзенрот М. П., Фейзал А. А., Чен С. Математика в машинном обучении. – М.: ДМК Пресс, 2024. – 512 с.
7. Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. – М.: ДМК Пресс, 2016. – 302 с.
8. Воронцов К. В. Машинное обучение: курс лекций [Электронный ресурс]. — Режим доступа — URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов) (Дата обращения 17.06.2026)