КОТЛОВСКИЙ М.Ю.,
д.м.н., ФГБУ «ЦНИИОИЗ» Минздрава России, Москва, Россия,
e-mail: m.u.kotlovskiy@mail.ru

ЦЫБИКОВА Э.Б.,
д.м.н., ФГБУ «ЦНИИОИЗ» Минздрава России, Москва, Россия,
e-mail: erzheny2014@yandex.ru

ЛОРСАНОВ С.М.,
Министерство здравоохранения Чеченской Республики, г. Грозный, Россия,
e-mail: info@minzdravchr.ru

ФАДЕЕВ П.А.,
к.м.н., Министерство здравоохранения Чеченской Республики, г. Грозный, Россия,
e-mail: fadeipavel@mail.ru

ФАДЕЕВА С.О.,
Республиканский центр общественного здоровья и медицинской профилактики»,
г. Грозный, Россия; Ярославский государственный медицинский университет,
г. Ярославль, Россия, e-mail: fadeeva-lana@inbox.ru

ГУСЕВ А.В.,
к.т.н., ФГБУ «ЦНИИОИЗ» Минздрава России, Москва, Россия, e-mail: agusev@webiomed.ai

РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЧИСЛА ВПЕРВЫЕ ВЫЯВЛЕННЫХ ПАЦИЕНТОВ С ВИЧ ИНФЕКЦИЕЙ В СУБЪЕКТАХ РОССИЙСКОЙ ФЕДЕРАЦИИ

DOI: 10.25881/18110193_2023_3_16

Аннотация. *Цель: разработать модель прогнозирования числа впервые выявленных пациентов с ВИЧ-инфекцией в субъектах Российской Федерации с использованием методов машинного обучения*

Материалы и методы: исходные данные были получены из формы федерального статистического наблюдения №61 и данных Росстата о среднегодовой численности населения - из 85 субъектов РФ (2016-2022 годы). Проведено сравнение методов машинного обучения и их ансамблей при построении регрессионной модели для прогнозирования числа впервые выявленных пациентов с ВИЧ-инфекцией в субъектах РФ.

Результаты: модель строилась с помощью методов: линейной регрессии, решающего дерева, случайного леса, градиентного бустинга на решающих деревьях и бэггинга. Использовалась интерактивная вычислительная среда «Jupiter Notebook» (6.5.2) и программные библиотеки «Pandas» (1.5.3), «Scikit-learn» (1.0.2), «Statsmodels» (0.13.5) и CatBoost. Оптимальные гиперпараметры подбирались с использованием фреймворка «Optuna». В качестве метрик качества выступили: корень из среднеквадратичной ошибки (RMSE); коэффициент детерминации (R2); средняя абсолютная ошибка (MAE); средняя абсолютная процентная ошибка (MAPE); медианная абсолютная ошибка (MedAE).

Выводы: применение методов и алгоритмов машинного обучения дает разные результаты в части метрик точности работы моделей. Наихудшие значения всех метрик качества продемонстрировал метод линейной регрессии (MAPE 67%). Наилучшим являлось сочетание (Бэггинг) двух ансамблевых методов — случайного леса и градиентного бустинга на решающих деревьях, поскольку было достигнуто максимальное значение большего числа метрик качества. В этой связи целесообразно проверять все доступные методы и алгоритмы машинного обучения и затем выбирать из полученных результатов наиболее качественную модель.

Ключевые слова: *ВИЧ-инфекция, прогнозная аналитика, машинное обучение, искусственный интеллект.*

Для цитирования: *Котловский М.Ю., Цыбикова Э.Б., Лорсанов С.М., Фадеев П.А., Фадеева С.О., Гусев А.В. Разработка модели машинного обучения для прогнозирования числа впервые выявленных пациентов с ВИЧ инфекцией в субъектах Российской Федерации. Врач и информационные технологии. 2023; 3: 16-29. doi: 10.25881/18110193_2023_3_16.*

KOTLOVSKIY M.YU.,
DSc, Central Research Institute of Organization and Informatization of Healthcare of the Ministry of Health of Russia, Moscow, Russia, e-mail: m.u.kotlovskiy@mail.ru

TSYBIKOVA E.B.,
DSc, Central Research Institute of Organization and Informatization of Healthcare» Ministry of Health of Russia, Moscow, Russia, e-mail: erzheny2014@yandex.ru

LORSANOV S.M.,
Ministry of Health of the Chechen Republic, Grozny, Russia,
e-mail: info@minzdravchr.ru

FADEEV P.A.,
PhD, Ministry of Health of the Chechen Republic, Grozny, Russia,
e-mail: fadeipavel@mail.ru

FADEEVA S.O.,
Republican Center for Public Health and Medical Prevention, Grozny, Russia;
Yaroslavl State Medical University, Yaroslavl, Russia, e-mail: fadeeva-lana@inbox.ru

GUSEV A.V.,
PhD, Central Research Institute of Organization and Informatization of Healthcare of the Ministry of Health of Russia, Moscow, Russia, e-mail: agusev@webiomed.ai

DEVELOPMENT OF A MACHINE LEARNING MODEL PREDICTING THE INCIDENCE OF NEWLY DIAGNOSED HIV INFECTION IN THE SUBJECTS OF THE RUSSIAN FEDERATION

DOI: 10.25881/18110193_2023_3_16

Abstract. *Aim: to develop a model predicting the incidence of newly diagnosed HIV infection in the subjects of the Russian Federation using machine learning methods.*

Materials and methods: The initial data were obtained from the Federal statistical observation Form No. 61 and Rosstat data on the average annual population of 85 subjects of the Russian Federation (2016-2022). We made a comparison of machine learning methods and their ensembles in the construction of a regression model for predicting the incidence of newly diagnosed patients with HIV infection in the subjects of the Russian Federation.

Results: The model was built using the following methods: linear regression, decision Tree, random forest, gradient boosting on decision trees (GBDT) and bagging. The interactive computing environment «Jupiter Notebook» (6.5.2) and software libraries «Pandas» (1.5.3), «Scikit-learn» (1.0.2), «Statsmodels» (0.13.5) and CatBoost were utilized. Optimal hyperparameters were selected using the Optuna framework. The following quality metrics were used: root of mean square error (RMSE); coefficient of determination (R2); average absolute error (MAE); average absolute percentage error (MAPE); median absolute error (MedAE).

Conclusions: The use of machine learning methods and algorithms gives different results in terms of metrics of model accuracy. The worst values of all quality metrics were demonstrated by the linear regression method (MAPE 67%). The combination (bagging) of the two ensemble methods — Random Forest and GBDT — was the best, since the highest values were found for a larger number of quality metrics. In this regard, it is reasonable to test all available machine learning methods and algorithms and then select the best-quality model from the results obtained.

Keywords: *HIV infection, predictive analytics, machine learning, artificial intelligence.*

For citation: *Kotlovsky M.Yu., Tsybikova E.B., Lorsanov S.M., Fadeev P.A., Fadeeva S.O., Gusev A.V. Development of a machine learning model predicting the incidence of newly diagnosed HIV infection in the subjects of the Russian Federation. Medical doctor and information technology. 2023; 3: 16-29. doi: 10.25881/18110193_2023_3_16.*

ВВЕДЕНИЕ

Профилактика ВИЧ-инфекции среди населения подразделяется на первичную, направленную на исключение вероятности проникновения вирусных частиц в организм человека, и вторичную, при которой проводится профилактика и лечение заболеваний, способствующих заражению ВИЧ-инфекцией [1]. Одним из основных индикаторов, свидетельствующих об эффективности проводимых профилактических мер в субъектах Российской Федерации (РФ), является число впервые выявленных пациентов с ВИЧ-инфекцией [1]. Разработка методов моделирования для принятия управленческих решений, направленных на повышение эффективности проводимой первичной и вторичной профилактики ВИЧ-инфекции, является важной научно-практической задачей, в решении которой большую поддержку может оказать использование методов машинного обучения [2]. В настоящее время в ряде публикаций российских и зарубежных авторов представлены данные об использовании методов машинного обучения при построении прогнозных моделей для различных целевых событий, таких как оценка продолжительности жизни пациентов с ВИЧ-инфекцией, прогнозирование госпитальной летальности, оценка распространенности резистентности к антиретровирусным препаратам среди пациентов с ВИЧ-инфекцией и оценка риска заболеваемости ВИЧ-инфекцией [3–8]. Вместе с тем все еще остаются малоизученными вопросы, посвященные применению машинного обучения для создания моделей, позволяющих прогнозировать число впервые выявленных пациентов ВИЧ-инфекцией.

Цель исследования: разработать модель прогнозирования числа впервые выявленных пациентов с ВИЧ-инфекцией в субъектах Российской Федерации с использованием методов машинного обучения.

МАТЕРИАЛЫ И МЕТОДЫ

Для проведения исследования использованы данные из формы федерального статистического наблюдения №61 по 85 субъектам РФ за 2016–2022 годы и данные Росстата о среднегодовой численности населения данных субъектов РФ. Для построения прогностической модели были использованы следующие данные 1) среднегодовая численность населения субъектов РФ (Насел); 2) число лиц, обследованных с использованием методов лабораторной диагностики для выявления ВИЧ-инфекции в субъектах РФ (Блоттинг); 3) контингенты с ВИЧ-инфекцией, состоявшие под диспансерным наблюдением в СПИД-центрах в субъектах РФ на конец календарного года (Конт_ВИЧ).

Были рассчитаны следующие показатели: 1) Распространённость ВИЧ-инфекции

$$(\text{Распростр.} = \frac{\text{Конт. ВИЧ}}{\text{Насел.}} \text{ X } 100000);$$

2) Охват населения лабораторным обследованием для выявления ВИЧ–инфекции

$$(\text{ИФА\_иссл.} = \frac{(\text{ИФА\_иссл.})}{\text{Насел.}} \text{ X } 100000).$$

В качестве целевой переменной, используемой для прогноза при применении методов машинного обучения, явилось число впервые выявленных пациентов с ВИЧ-инфекцией в субъектах РФ.

Для работы с данными, полученными из 85 субъектов РФ, использовалась интерактивная вычислительная среда «Jupiter Notebook» (6.5.2) [9]. Обработка и анализ данных, а также составление и работа со структурированным датасетом производилась в программной библиотеке «Pandas» (1.5.3) [10]. Для построения линии тренда и нахождения коэффициентов линейной регрессии применялись программные библиотеки «Scikit-learn» (1.0.2) и «Statsmodels» (0.13.5) [11, 12].

Для построения прогностических моделей использованы методы линейной регрессии (Linear Regression), решающего дерева (Decision Tree), случайного леса (Random Forest), градиентного бустинга на решающих деревьях (GBDT), бэггинга и программные библиотеки Scikit-learn и CatBoost [13].

Оптимальные гиперпараметры прогностических моделей подбирались с использованием фреймворка «Optuna» [14].

В качестве метрик качества построенных моделей выступили: корень из среднеквадратичной ошибки (RMSE); коэффициент детерминации (R2); средняя абсолютная ошибка (MAE); средняя абсолютная процентная ошибка (MAPE); медианная абсолютная ошибка (MedAE). Для расчета

данных показателей использованы алгоритмы из модуля «Metrics» библиотеки «Scikit-learn» [15].

РЕЗУЛЬТАТЫ

В данном исследовании изначально было установлено, что значения **целевой объясняемой переменной** (число впервые выявленных пациентов с ВИЧ-инфекцией), значения которой мы прогнозировали, имели интервальную шкалу измерения. В связи с этим, методы машинного обучения, использованные для построения модели, относились к категории **регрессионных методов обучения с учителем**.

Целевая объясняемая переменная не имела нормального распределения. Это было подтверждено тестами Колмогорова-Смирнова и Шапиро-Уилка ($p<0,05$). Распределение данной переменной по своей форме напоминало распределение Пуассона (рис. 1), что накладывало ограничение на применение ряда методов и могло снижать прогностическую точность построенных моделей. В связи с этим для улучшения качества прогноза целевая объясняемая переменная была преобразована путем нахождения натурального логарифма каждого из ее значений, что приближало ее распределение к нормальному. В свою очередь, после подбора оптимальных гиперпараметров моделей, исходные параметры которых задает сам исследователь, и получения их предсказаний с использованием тестовых данных, для нахождения значений итоговых метрик качества проводилось обратное преобразование путем нахождения экспоненты каждого члена как преобразованных фактических данных, так и полученных предсказаний.

Первоначально была исследована предиктивная способность линейных и нелинейных одиночных методов машинного обучения и построенных на их основании моделей. Данными методами явились линейная регрессия и решающее дерево для регрессии [16–19].

Первая предиктивная модель была построена методом линейной регрессии, являющимся наиболее простым и изученным линейным методом машинного обучения, в котором предсказанные значения объясняемой целевой переменной y определяются через нахождение свободного члена b_0, и коэффициентов $b_{1...k}$ для каждой из объясняющих переменных $X_{1...k}$, на основании которых делается прогноз в виде линии тренда. При этом математический алгоритм выстраивает прямую линию, максимально приближенную ко всем реальным значениям целевой объясняемой переменной $f(x, b) = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_k * X_k$.

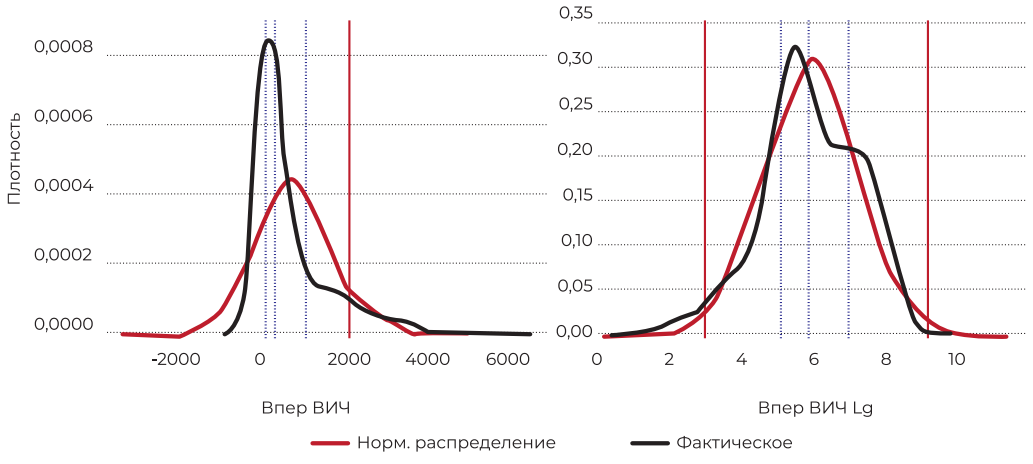

Рисунок 1 — Распределение значений целевой объясняемой переменной, 85 субъектов Российской Федерации, 2016–2022 годы, абсолютные значения. Здесь и далее: пунктирной вертикальной линией обозначены (квартили Q1, Q2, Q3) соответственно, сплошной вертикальной линией обозначены пределы доверительного интервала.

Данный метод является простым и интуитивно понятным. В нем отсутствует большое число гиперпараметров для настройки. Однако он демонстрирует хорошие результаты при наличии линейной зависимости между целевой и объясняющими переменными. При использовании данного метода для построения предиктивных моделей (но не для исследования связей) этот недостаток может быть преодолён за счет генерации дополнительных признаков путем проведения степенных преобразований значений объясняющих переменных. Кроме того, к недостаткам данного метода относятся: повышенная чувствительность к разной масштабности показателей объясняющих переменных при нарушении их нормального распределения, наличие «выскакивающих» значений и сильно связанных между собой переменных [20].

В проведенном исследовании было установлено, что каждая из объясняющих переменных, на основе которых строился прогноз, как и объясняемая переменная, имели ненормальное

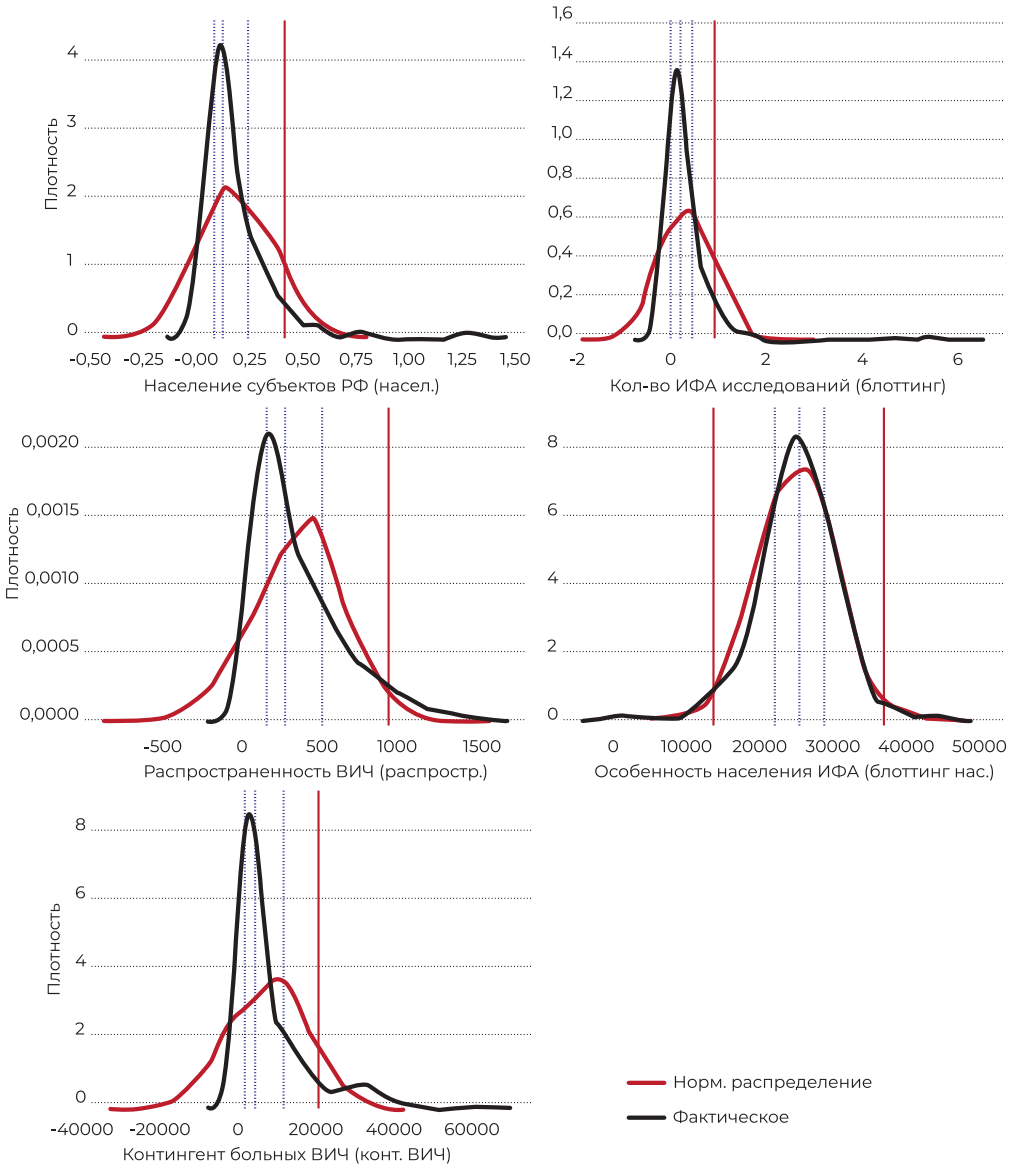

Рисунок 2 — Распределение значений объясняющих переменных, 85 субъектов Российской федерации, 2016–2022 годы, абсолютные значения.

распределение из-за наличия выскакивающих значений (рис. 2). Это было подтверждено математически путем нахождения значений тестов Колмогорова-Смирнова и Шапиро-Уилка ($p<0,05$). В дополнение к этому значения переменных лежали в разных числовых диапазонах.

В связи с вышеперечисленными особенностями используемых данных для применения метода линейной регрессии (и других методов), была проведена их предподготовка. Для этого было произведено робастное шкалирование. Для этого использовали алгоритм RobustScaler из программного модуля Preprocessing бесплатной библиотеки машинного обучения для языка программирования Python-Scikit-learn [21]. При этом из каждого значения удалялась медиана (определенная на тренинговых данных), и происходило масштабирование значения в соответствии с интерквартильным размахом (тренинговых данных), диапазоном между 1-м квартилем (25-й квартиль) и 3-м квартилем (75-й квартиль).

Данное шкалирование применялось и в работе с остальными методами машинного обучения. Поскольку переменные закономерно сильно коррелировали между собой, применяли дополнительно L1 регуляризацию (Lasso — регрессию), в основе которой лежит идея добавления штрафного коэффициента к первоначальной функции потерь, что позволяет произвести разряжение и из всего массива объясняющих переменных отобрать наиважнейшие, задающие тенденцию, удалив все остальное (шум) [22]. В результате из модели в автоматическом режиме были исключены коррелирующие переменные и переменные с низкой предсказательной ценностью. Для этого была использована Lasso модель из модуля — «Linear_model» библиотеки «Scikit-learn» [18].

Далее был проведен поиск оптимального значения единственного гиперпараметра — α. Это значение штрафного коэффициента для функции потерь, которое по умолчанию было равно 1. Подбор значений производился в диапазоне от 0,001 до 20. Для подбора оптимального значения гиперпараметров здесь и далее применяли фреймворк Optuna, предназначенный для автоматизированного поиска оптимальных значений гиперпараметров методов машинного обучения [14]. В качестве функции потерь была использована функция вычисления среднеквадратичной ошибки (MSE), конкретно в данном случае с L1 регуляризацией. Предиктивная эффективность каждой модели проверялась методом кросс валидации на 5 фолдах (блоках) путем максимизации среднего значения негативной среднеквадратичной логарифмической ошибки (MSLE) [23]. Благодаря большому «гасящему» эффекту логарифма, данная функция потерь более применима к данным, имеющим разброс значений в несколько порядков, что имело место и в нашем случае. Для нахождения наилучшего значения штрафного коэффициента было проведено 10000 итераций (пробных построений модели). **Это заняло 16 минут**. Наилучшее найденное значение гиперпараметра α было установлено на 3839 итерации и составило 0,001, то есть наилучшие значения предсказаний были получены при штрафном коэффициенте, стремящемся к нулю, или полном отсутствии регуляризации.

Каждый анализируемый признак имел свою предсказательную ценность (рис. 3). Среди них основной ценностью обладали такие признаки, как численность населения (37%) и распространённость ВИЧ-инфекции (33%). Далее следовало число проведенных обследований (16%) и численность контингентов с ВИЧ-инфекцией (13%). Предсказательная ценность практически отсутствовала у признака — обеспеченность населения лабораторным обследованием для выявления ВИЧ-инфекции (1%).

В построенном уравнении регрессии свободный член был равен 5,52. Установленные коэффициенты были положительными для следующих признаков: численность населения — 1,51; обеспеченность населения лабораторным обследованием — 0,05; распространённость ВИЧ-инфекции — 1,37. В тоже время значения ряда коэффициентов были отрицательными для таких признаков, как число проведенных обследований -0,65 и численность контингентов с ВИЧ-инфекцией -0,54, что могло быть обусловлено наличием сильных корреляционных связей. Удаление одной из коррелирующих переменных снижало предсказательную ценность модели, как и построение новых полиномиальных и обобщенных моделей на основе имеющихся переменных.

После подбора оптимальных гиперпараметров модели выборку разделяли на обучающую

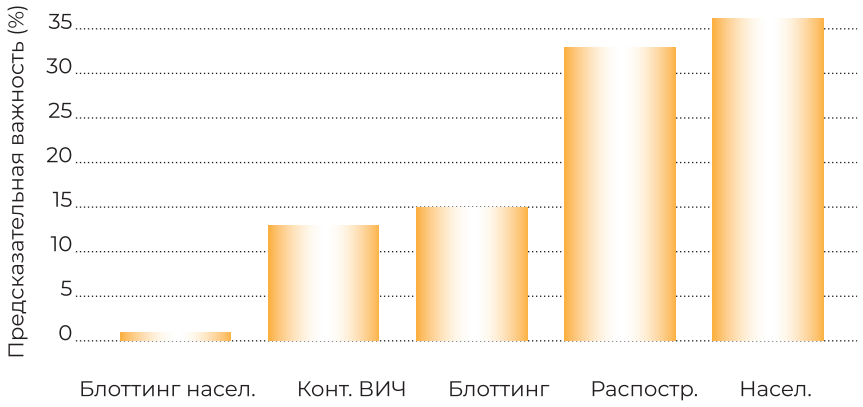

Рисунок 3 — Предсказательная ценность переменных при построении линейной регрессионной модели (с L1 регуляризацией).

и тестовую в соотношении 9 к 1. Производили обучение модели и делали прогноз. После этого вычисляли значение метрик качества, используя экспоненту предсказанных и фактических значений (операция обратная ранее проведенному логарифмированию).

Вторым примененным методом машинного обучения было решающее дерево [19]. Это алгоритм используется для построения одиночной и, в данном случае, нелинейной предиктивной модели. Устройство данного дерева включает в себя так называемые «ветви» и «листья». На «ветвях» решающего решения записаны признаки, от которых зависит значение целевой переменной, а в «листьях», которыми заканчиваются «ветви», записаны ее значения. Чтобы сделать прогноз, необходимо опуститься по дереву вплоть до листа и получить соответствующее значение. Метод отличает простота и высокая (не уступающая линейной регрессии) скорость построения модели, а также интуитивная понятность и наличие возможности графического отображения. Он менее требователен к наличию линейной связи, нормальности распределения, отсутствию выбросов, одинаковому масштабу данных и отсутствию корреляций.

Для данного метода использовалась базовая модель «DecisionTreeRegressor» из модуля «Tree» библиотеки «Scikit-learn» и производился подбор таких гиперпараметров как:

- глубина решающего дерева (max_depth) в диапазоне от 2 до 30;
- минимальное число наблюдений, необходимое для разделения внутреннего узла (min_samples_split), в диапазоне от 1 до 50;
- минимальное число наблюдений, необходимых для образования листа (min_samples_leaf), в диапазоне от 1 до 50 [19].

Всего было проведено 10000 итераций в течение 34 минут. Наилучшее сочетание гиперпараметров было установлено на 5056 итерации.

Несмотря на то, что данная модель не требовала нормального распределения и была устойчива к наличию выбросов, применяли масштабирование используя RobustScaler [21]. В результате подобранная модель с наилучшими показателями имела следующие характеристики: 1) max_depth — 7; 2) min_samples_split — 25; 3) min_samples_leaf — 5.

При этом каждая объясняющая переменная имела свою прогностическую ценность, основной из которых обладала такая переменная, как численность контингентов с ВИЧ-инфекцией (94%). Более низкую значимость показала переменная — численность населения (4%). Прогностическая ценность других переменных имела следовые значения (охват обследованием -0,9%, распространенность ВИЧ-инфекции -0,7%, обеспеченность населения лабораторным обследованием -0,4%).

Далее была изучена предиктивная способность моделей, построенных на основании более сложных нелинейных ансамблевых методов и их сочетаний. **Третья** предиктивная модель была построена на основании

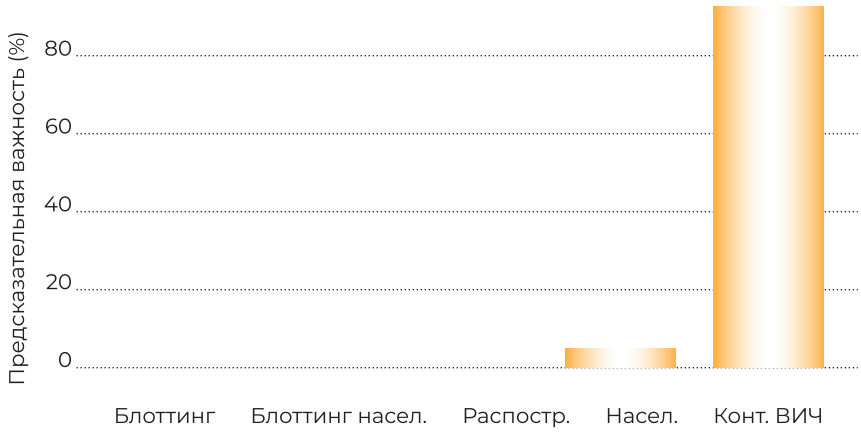

Рисунок 4 — Предсказательная ценность переменных при построении модели «решающего дерева», 85 субъектов Российской Федерации, 2016–2022 годы, абсолютные значения.

алгоритма случайного леса [24]. При этом основной идеей является использование большого количества «решающих деревьев», каждое из которых в отдельности даёт относительно невысокое качество прогноза, но их общее число делает суммарный результат более высоким. Особенностью этого метода является смещение значения полученного прогноза при низком разбросе, вызванном переобучением. Кроме того, каждое «решающее дерево» достраивается независимо от другого, что дает возможность распараллелить вычисления.

Для данного метода использовалась модель «RandomForestRegressor» из модуля «Ensemble» библиотеки «Scikit-learn» [25]. Данные подвергались робастному шкалированию. Производился подбор оптимальных значений следующих гиперпараметров:

- параметр сложности, используемый для снижения сложности с минимальными затратами (ccp_alpha), в диапазоне от 0,0001 до 5;
- максимальная глубина дерева (max_depth) от 2 до 30;
- количество деревьев в ансамбле (n_estimators) от 1 до 10000;
- минимальное число наблюдений, необходимое для разделения внутреннего узла (min_samples_split), от 2 до 50;
- минимальное число наблюдений, необходимых для образования листа (min_samples_leaf), от 1 до 30;
- минимальная величина примесей (min_impurity_decrease) от 0,001 до 5 (log = True).

Было проведено 10000 итераций в течение 2 часов 49 минут. Наилучшее сочетание гиперпараметров было установлено на 5056 итерации.

В результате подобранная модель с лучшими показателями имела следующие характеристики: 1) ccp_alpha — 0,0001; 2) max_depth — 8; 3) n_estimators — 2112; 4) min_samples_split — 3; 5) min_samples_leaf -2; 6) min_impurity_decrease 0,001.

Было установлено, что, как и в случае построения простого «решающего дерева», основной предсказательной ценностью обладала переменная — контингенты с ВИЧ-инфекцией (91%) (рис. 5). Низкую ценность показали переменные — численность населения (5%), число обследованных лиц (2%), распространённость ВИЧ-инфекции (1%) и обеспеченность населения лабораторным обследованием (1%).

На следующем шаге для построения предиктивной модели использован более сложный алгоритм машинного обучения, построенный на принципах градиентного бустинга над решающими деревьями [26]. Бустинг — это техника ансамблей методов машинного обучения, в основе которой лежит последовательное обучение нескольких моделей для повышения точности всей системы [13]. Этот метод использует идею о том, что каждая последующая модель будет учиться на ошибках предыдущей. Таким

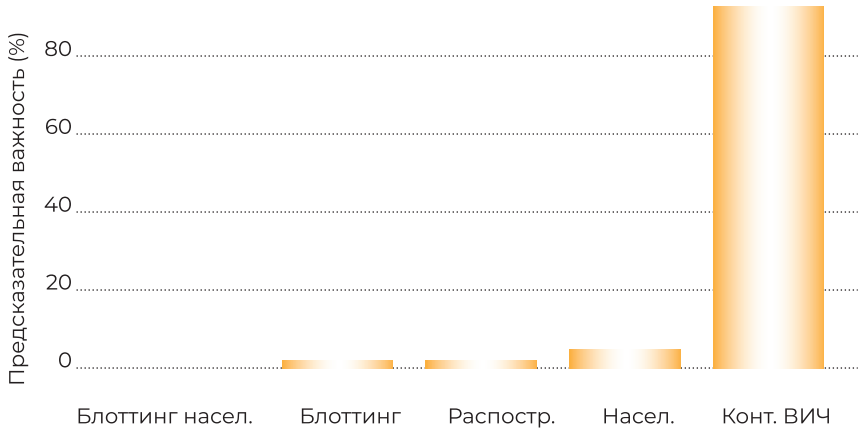

Рисунок 5 — Предсказательная важность переменных при построении предиктивной модели «случайный лес», 85 субъектов Российской Федерации, 2016–2022 годы, абсолютные значения.

образом, построение «решающих деревьев» в нашем случае происходит последовательно одно за другим, что требует гораздо больших временных затрат, чем предыдущие методы. Также особенностями данного метода, с одной стороны, является низкое смещение предсказанного значения, а с другой — склонность к переобучению и высокому значению разброса предсказаний.

Нами использована отечественная разновидность метода машинного обучения — «CatBoostRegressor» из программной библиотеки «CatBoost» [13]. Произведен подбор оптимальных значений следующих гиперпараметров модели:

- штрафного коэффициента (‘penalties_coefficient) в диапазоне от 0.1 до 10;
- максимальной глубины дерева (max_depth) от 2 до 16;
- количества деревьев в ансамбле (num_trees) от 1 до 10000;
- минимального число наблюдений, необходимых для образования листа (min_data_in_leaf), от 1 до 30.

Было проведено 100 итераций в течение 13 часов 30 мин. Наилучшие показатели были получены на 78 итерации.

В результате подобранная модель с наилучшими показателями имела следующие характеристики: 1) penalties_coefficient — 5,56; 2) max_depth — 5; 3) num_trees — 3224; 4) min_data_in_leaf — 6.

Было установлено, что наибольшей предсказательной ценностью обладала переменная — контингенты с ВИЧ-инфекцией (44%) (рис. 6).

Далее в порядке убывания следовали: численность населения (24%), число обследованных лиц (14%), распространённость ВИЧ-инфекции (11%) и обеспеченность населения лабораторным обследованием (7%). При этом полученные значения каждой из используемых метрик качества, за исключением медианной ошибки, превышали таковые у предшествующих моделей.

Следующим использованным методом стал Бэггинг, суть которого заключается в комбинации предсказанных значений независимых методов [27]. В работе был произведен расчет предсказательной способности обоих проанализированных ансамблевых методов — случайного леса и градиентного бустинга на решающих деревьях. При этом производили сложение предсказанных значений в подобранных оптимальных соотношениях y = 0,3RandomForest + 0,7 GBDT.

ОБСУЖДЕНИЕ

Полученные значения метрик качества каждой из построенных прогностических моделей представлены в таблице 1. При этом мы

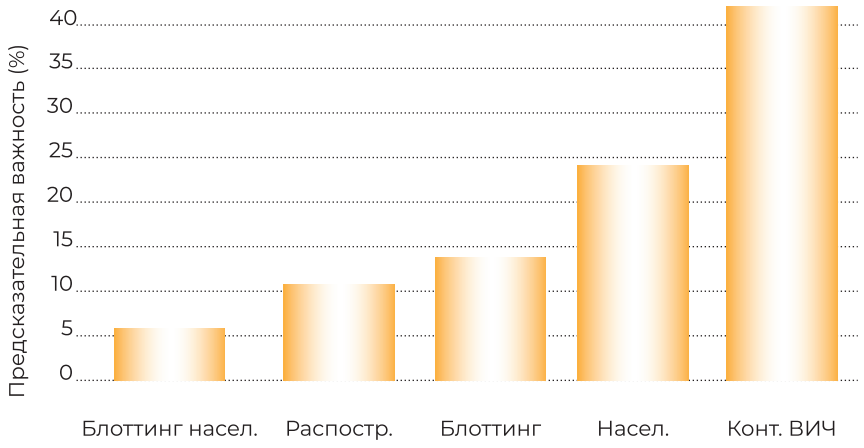

Рисунок 6 — Распределение переменных в зависимости от их прогностической ценности при построении предиктивной модели «градиентный бустинг на решающих деревьях», 85 субъектов Российской Федерации, 2016–2022 годы, абсолютные значения.

Таблица 1 — Метрики качества построенных предиктивных регрессионных моделей

Метрика качества	Линейная регрессия + (L1 – рег.)	Решающее дерево	Случайный лес	градиентный бустинг на решающих деревьях	Случайный лес + градиентный бустинг на решающих деревьях
R2 ,%	27,51	76,96	88,62	90,74	94,31
RMSE абс. число	827,46	466,57	327,85	295,69	293,91
MAE абс. число	349,92	274,77	213,25	190,25	192,48
MAPE, %	67	38	30,79	30,45	29,97
MedAE абс. число	119,18	99,78	90,67	103,34	114,71

старались получить наибольшее максимально приближенное к 100 % значение R2 и наименьшие значения остальных показателей, говорящих об абсолютном (RMSE, MAE, MedAE) или относительном (MAPE) значении величины ошибки.

Используемые метрики качества были подобраны на основании интуитивной понятливости. При этом каждая из них могла иметь свои недостатки, поэтому прогностическая ценность построенных моделей оценивалась комплексно. Наихудшие значения всех метрик качества продемонстрировал метод линейной регрессии (MAPE 67%). Наилучшим было сочетание (Бэггинг) двух ансамблевых методов: случайного леса, градиентного бустинга на решающих деревьях, поскольку было установлено наилучшее значение большего числа метрик качества (R^2, RMSE, MAPE). У ансамблевого метода случайный лес наилучшим среди остальных было значение метрики качества — MedAE. Хорошо проявила себя отечественная разновидность метода градиентного бустинга на решающих деревьях — алгоритм CatBoost, наилучшим было значение показателя MAE. При этом данный метод показал лучшие, чем случайный лес, значения четырех из пяти метрик качества (R^2, RMSE, MAPE, MAPE). Однако конструктивные особенности применяемого метода требовали значительного расхода времени для подбора оптимальных

гиперпараметров и обучения модели. Также требовалась высокая производительность ЭВМ. Это может стать преградой для включения дополнительных объясняющих переменных и увеличения числа наблюдений при построении модели. В тоже время применение виртуальных машин таких сервисов, как «Google Colab», «Kaggle», «Paperspace Gradient», «Deepnote», «Yandex DataSphere» и др., для повышения вычислительных возможностей может быть ограничено лишь характером анализируемых данных.

ВЫВОДЫ

Результаты проведенного исследования показали, что применение современных методов и алгоритмов машинного обучения может давать различные результаты в части метрик точности работы моделей, поэтому при решении прикладных задач целесообразно проверять все доступные методы и алгоритмы машинного обучения и затем выбирать из полученных результатов наиболее качественную модель. Такой подход требует большие затраты времени на подготовку данных, проведение экспериментов с различными вариантами машинного обучения, а затем дополнительную настройку и поиск гиперпараметров в выбранной версии модели. В этой связи целесообразно проводить дальнейшие исследования в части применения технологий автоматизированного машинного обучения (AutoML), а также применения высокопроизводительных вычислительных комплексов и специализированных платформ, ускоряющих процессы обучения.

ЛИТЕРАТУРА/REFERENCES

1. ВИЧ-инфекция и СПИД. Национальное руководство. Под ред. акад. РАН, профессора В.В. Покровского, Москва: ГЭОТАР-МЕДИА, 2020. — 686 с. [HIV infection and AIDS. National leadership. Acad. RAS, Professor V.V. Pokrovsky, editor. Moscow: GEOTAR-MEDIA, 2020. 686 p. (In Russ.)]
2. Бодрин К.А., Красноперова А.А. Использование технологий машинного обучения в медицине // Теория и практика современной науки. — 2018. — №10(40). — С.52-56. [Bodrin KA, Krasnoperova AA. The use of machine learning technologies in medicine. Theory and practice ofmodern science. 2018; 10(40): 52-56. (In Russ.)]
3. Востроknутов М.Е., Дюжева Е.В., Кузнецова А.В., Сенько О.В. Факторы риска госпитальной летальности больных с сочетанием туберкулеза и ВИЧ-инфекции в учреждениях уголовно-исполнительной системы // Туберкулез и болезни легких. — 2019. — Т.97. — №7. — С.34-41. [Vostroknutov ME, Dyuzheva EV, Kuznetsova AV, Senko OV. Risk factors of hospital mortality of patients with a combination of tuberculosis and HIV infection in institutions of the penal system. Tuberculosis and lung diseases. 2019; 97(7): 34-41. (In Russ.)] doi: 10.21292/2075-1230-2019-97-7-34-41.
4. Тарасова О.А., Филимонов Д.А., Пороikов В.В. Компьютерный прогноз резистентности вируса иммунодефицита человека к ингибиторам обратной транскриптазы ВИЧ // Биомедицинская химия. — 2017. — Т.63. — №5. — С.457-460. [Tarasova OA, Filimonov DA, Poroikov VV. Computer prediction of human immunodeficiency virus resistance to HIV reverse transcriptase inhibitors. Biomedical chemistry. 2017. 63(5): 457-460. (In Russ.)] doi: 10.18097/PBMC20176305457.
5. Rajendran M, Ferran MC, Mouli L, Babbitt GA. Lynch Evolution of drug resistance drives destabilization of flap region dynamics in HIV-1 protease. Biophys Rep (NY). 2023; 3(3): 100121. doi: 10.1016/j.bpr.2023.100121.
6. Bukic E, Milasin J, Toljic B, Jadzic J, Jevtovic D, Obradovic B, Dragovic G. Association between Combination Antiretroviral Therapy and Telomere Length in People Living with Human Immunodeficiency Virus. Biology (Basel). 2023; 12(9): 1210. doi: 10.3390/biology12091210.
7. Birri Makota RB, Musenge E. Predicting HIV infection in the decade (2005-2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. PLOS Digit Health. 2023; 2(6): e0000260. doi: 10.1371/journal.pdig.0000260.
8. Mamo DN, Yilma TM, Fekadie M, Sebastian Y, Bizuayehu T, Melaku MS, Walle AD. Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. BMC Med Inform Decis Mak. 2023; 23(1): 75. doi: 10.1186/s12911-023-02167-7.

9. Jupyter Notebook. Available at: https://docs.jupyter.org/en/latest/. Accessed 10.10.2023.
10. Pandas. Available at: https://pandas.pydata.org/docs/. Accessed 10.10.2023.
11. Scikit-learn. Documentation. Available at: https://scikit-learn.org/stable/index.html. Accessed 10.10.2023.
12. Statsmodels. Available at: https://www.statsmodels.org/stable/user-guide.html. Accessed 10.10.2023.
13. CatBoost. Available at: https://catboost.ai/en/docs/. Accessed 10.10.2023.
14. Optuna. Available at: https://optuna.org/#key_features. Accessed 10.10.2023.
15. Scikit-learn. Evaluation of models. Available at: https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed 10.10.2023.
16. Лысенко А.А. Введение в регрессионный анализ данных и регрессионные модели // Труды Санкт-Петербургского государственного морского технического университета. — 2020. — Т.1. — №S2. — С.15. [Lysenko AA. Introduction to regression analysis of data and regression models. Proceedings of the St. Petersburg State Maritime Technical University. 2020; 1(S2): 15. (In Russ.)]
17. Пернебай Б.А. Python. Регрессия дерева решений с использованием sklearn // Polish Journal of Science. — 2021. — №38-1(38). — С.51-56. [Pernebai BA. Python. Decision tree regression using sklearn. Polish Journal of Science. 2021; 38-1(38): 51-56. (In Russ.)]
18. Scikit-learn. Linear models. Available at: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model. Accessed 10.10.2023.
19. Scikit-learn. Decision tree, regressor. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor. Accessed 10.10.2023.
20. Scikit-learn. Common errors in the interpretation of linear model coefficients. Available at: https://scikit-learn.org/stable/auto_examples/inspection/plot_linear_model_coefficient_interpretation.html#sphx-glr-auto-examples-inspection-plot-linear-model-coefficient-interpretation-py. Accessed 10.10.2023.
21. Scikit-learn. Robust scaling. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html. Accessed 10.10.2023.
22. Scikit-learn. Lasso regression. Available at: scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html. Accessed 10.10.2023.
23. Scikit-learn. Cross-validation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html. Accessed 10.10.2023.
24. Носова Г.С., Абдуллин А.Х. Машинное обучение на основе непараметрического и нелинейного алгоритма Random Forest (RF) // Инновации. Наука. Образование. — 2021. — №35. — С.33-39. [Nosova GS, Abdullin AH. Machine learning based on nonparametric and nonlinear Random Forest (RF) algorithm. Innovation. The science. Education. 2021; 35: 33-39. (In Russ.)]
25. Scikit-learn. Random forest, regressor. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble. RandomForestRegressor. html. Accessed 10.10.2023.
26. Zhang И, Ren J, Wei Z, et al. Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm. IEEE Access. 2019; 7: 32423-32433. doi: 10.1109/ACCESS.2019.2902217.
27. Plaia A, Buscemi S, Fürnkranz J, Mencía EL. Comparing Boosting and Bagging for Decision Trees of Rankings. Journal of Classification. 2022; 39(1): 78-99. doi: 10.1007/s00357-021-09397-2.