Сравнительный анализ точности модели автоматизированного машинного обучения для выявления сердечно-сосудистых заболеваний

Т.В. Афанасьева, А.П. Кузлякин, А.В. Комолов

***Аннотация* — Сердечно-сосудистые заболевания (ССЗ) широко распространены среди пациентов с хроническими неинфекционными заболеваниями и являются одной из ведущих причин смертности населения, в том числе в трудоспособном возрасте. Разработка пациент-ориентированных систем для раннего выявления сердечно-сосудистых заболеваний с использованием моделей машинного обучения является перспективным направлением, интегрирующим медицинские знания и информационные интеллектуальные технологии для систем поддержки принятия врачебных решений. Для упрощения и ускорения процесса разработки конкретного такого решения на основе большого разнообразия моделей машинного обучения активно развивается направление автоматического машинного обучения (AutoML). В статье приведен сравнительный анализ точности модели AutoML, создаваемой с использованием библиотеки AutoGluon-Tabular. Сравнение точности было реализовано в двух направлениях: по отношению к трем сценариям предобработки данных пациентов с ССЗ и по отношению к базовым моделям машинного обучения, содержащихся в библиотеке AutoGluon-Tabular. Сравнительный анализ на открытой базе данных UCI показал, что точность модели AutoML в выявлении сердечно-сосудистых заболеваний варьируется в диапазоне от 87,41% до 95,65%, причем максимальная точность получена в сценарии с Z-нормализацией исходных данных, а минимальная при использовании встроенного в AutoML алгоритма предобработки данных.**

***Ключевые слова* — сердечно-сосудистые заболевания, автоматизированное машинное обучение, классификация.**

I. Введение

В настоящее время многие страны сталкиваются с перегруженной системой здравоохранения и нехваткой квалифицированных врачей, поэтому большие перспективы показывают системы, ориентированные на пациента (ПОС), в которых активно используются модели машинного обучения (ML). Целью таких систем является предоставление людям возможности улучшить свое здоровье с помощью цифровых технологий, используемых для профилактики заболеваний на индивидуальном уровне и в домашних условиях.

Пациент-ориентированные системы включают в себя системы медицинских рекомендаций, виртуальных помощников, чат-ботов и мониторы симптомов, которые используют модели машинного обучения для оценки состояния здоровья, ранней диагностики заболеваний и прогнозирования вероятности серьезных событий, требующих госпитализации [1, 2, 3].

Выявление и диагностика ССЗ является актуальной задачей в связи с большим процентом фатальных исходов этого заболевания, в том числе у трудоспособного населения, и с увеличением продолжительности жизни населения. В публикации [4] описываются основы, исследовательские цели ML и актуальность для решения задач диагностики заболеваний в здравоохранении, а также в области медицины, где методы ML уже используются для анализа данных пациентов и ранней диагностики заболеваний, выявления признаков эпидемии или пандемии, а также для разработки лекарств.

Авторы статьи [5] сравнили точность моделей ML, таких как случайный лес, дерево решений, многослойный персептрон и модель бустинга деревьев решений XGBoost для диагностики и прогноза ССЗ. Модели были обучены на реальном наборе данных Kaggle из 70 000 экземпляров и достигли следующей точности: дерево решений: 86,37% (с перекрестной проверкой) и 86,53% (без перекрестной проверки), XGBoost: 86,87% (с перекрестной проверкой) и 87,02% (без перекрестной проверки), случайный лес: 87,05% (с перекрестной проверкой) и 86,92% (без перекрестной проверки), многослойный персептрон: 87,28% (с перекрестной проверкой) и 86,94% (без перекрестной проверки). Вывод, сделанный из этого исследования, заключается в том, что многослойный персептрон с с перекрестной проверкой превзошел все другие алгоритмы с точки зрения точности в 87,28%.

В статье [6] приведен систематический обзор моделей классификации для автоматизированной диагностики ССЗ на основе клинических признаков, изображений и ЭКГ, а также представлены некоторые направления

Статья получена 14 ноября 2023.
Т.В. Афанасьева, Российский экономический университет имени Г.В. Плеханова, Москва, Россия (e-mail: tv.afanasjeva@gmail.com)
А.П. Кузлякин, Российский экономический университет имени Г.В. Плеханова, Москва, Россия (e-mail: andrey-kuzliakin@yandex.ru)
А.В. Комолов, Российский экономический университет имени Г.В. Плеханова, Москва, Россия (e-mail: komolov_1995@mail.ru)

будущих исследований в области автоматического выявления заболеваний сердца на основе машинного обучения. Авторы рассмотрели некоторые ограничения ML-моделей в задаче классификации: обучение на большом объеме данных является сложной и трудоемкой задачей, модели машинного обучения могут страдать от проблемы переобучения, технология глубокого обучения требует подготовки и предобработки большого количества данных для обучения модели, что является дорогостоящей и сложной работой. Временная сложность – это еще одна проблема автоматического выявления сердечных заболеваний на основе подходов машинного обучения.

Понимая ограничения использования одной модели машинного обучения для выявления заболеваний сердца по результатам самоконтроля с помощью датчиков, авторы исследования [7] реализовали несколько моделей и предложил архитектуру веб-приложения для пациент-ориентированной системы, в которой автоматически выбирается лучшая ML модель по критерию точности. Исходными данными для выявления ССЗ были возраст, артериальное давление, сердцебиение, пол, результаты ЭКГ, уровень сахара в крови пациента. Модель показывает вероятность возникновения ССЗ.

Существующие проблемы в области применения ML в здравоохранении, в частости для выявления сердечно-сосудистых заболеваний (ССЗ), освещены в статье [8]. При этом отмечается, что в реальном мире существует относительно немного решений, поэтому присутствует значительный пробел в применении ML. Одой из причин указанной проблемы является достаточно трудоемкий процесс разработки конкретных решений ML, особенно в условиях, когда данные представлены в таблицах имеющихся медицинских информационных системах. Другой проблемой являются высокие требования к разработчикам систем ML в связи с необходимостью выбора подходящей модели из большого количества моделей ML и настройкой их параметров для получения релевантных показателей точности.

Одним из решений указанных проблем является использование моделей автоматизированного машинного обучения (AutoML), которые фокусируются на автоматизации процесса выбора модели, оптимальной настройке гиперпараметров и выделении признаков данных [9], в частности для табличных данных [10, 11].

В статье [12] представлен обзор публикаций в области AutoML для здравоохранения. Основным ограничением AutoML на данный момент по мнению авторов является то, что этот подход недостаточно эффективен для работы с большими данными, то есть вне исторических наборов данных малого и среднего размера. На основе анализа 101 статьи выявлены потенциальные возможности и препятствия для использования AutoML в здравоохранении, и показано, что эти автоматизированные методы могут улучшать производительность человека в определенных задачах машинного обучения.

Исследование [13] показало, что использование персептрона с гиперпараметрической настройкой позволяет с хорошей точностью реализовать бинарную классификацию пациентов на тех, у кого есть ССЗ, и тех, у кого их нет. В этом исследовании оптимизированы такие гиперпараметры, как ширина скрытого слоя, скорость обучения и функция активации с использованием метода случайного поиска по сетке. Разработанная модель искусственной нейронной сети на Кливлендском наборе данных, полученном из репозитория машинного обучения UCI, достигла точности в 93,44%, а также, было сокращено время, затрачиваемое на обучение модели.

Применение AutoML для выявления риска развития ССЗ в десятилетней перспективе с использованием данных о 423 604 участниках без ССЗ на начальном этапе в британском биобанке, приведено в статье [14]. Авторы сравнивают результаты модели AutoML с традиционным результатом оценки летальных исходов по шкале Фремингема. При использовании AutoML были получены наилучшие показатели точности: AUC-ROC: 0,774, 95% ДИ: 0,768-0,780.

Цель статьи состоит в разработке, применении и исследовании точности модели AutoML для выявления ССЗ пациентов, данные которых представлены в табличной форме.

Анализ публикаций в области применения моделей AutoML для выявления ССЗ показывает, что этот подход недостаточно представлен в публикациях, это затрудняет оценку его эффективности и ограничивает его применение для поддержки принятия врачебных решений в задаче диагностики ССЗ пациентов, данные которых содержатся в базах данных медицинских информационных системах. Поэтому результаты решения поставленной задачи в настоящей статье, а именно, задачи разработки и исследования модели AutoML по критерию точности классификации пациентов на тех, кто имеет и тех, кто не имеет ССЗ, позволит восполнить вышеобозначенный пробел.

Авторами были сформулированы следующие исследовательские вопросы:

1) Приведет ли использование модели AutoML, включающей оптимизированные базовые модели ML, к повышению точности выявления ССЗ по табличным данным пациентов по сравнению с базовыми моделями ML?
2) Как зависит точность модели AutoML от сценариев предобработки табличных данных?

II. Методология исследования

Для достижения поставленной цели и для решения исследовательских вопросов предлагается следующий подход, включающий следующие этапы:

1) Выбор и определение сценариев предварительной обработки открытых табличных данных для выявления ССЗ у пациентов.
2) Выбор модели AutoML и разработка с ее использованием системы классификации пациентов

на тех, кто имеет и тех, кто не имеет ССЗ.

3) Применение и сравнительная оценка моделей AutoML и базовых ML для выявления ССЗ на обучающей/тестовой выборках для различных сценариев предварительной обработки данных.

Под базовыми моделями машинного обучения понимаются те ML-модели, которые могут быть включены в состав результирующей AutoML модели из ее базы моделей. В данной статье использованы следующие оценки эффективности выявления ССЗ, традиционно применяемые в задачах классификации: accuracy (доля правильных ответов), recall (отношение правильно классифицированных объектов класса к общему количеству элементов этого класса), precision (доля правильно классифицированных объектов среди всех объектов, отнесенных к этому классу). Помимо рассматриваемых показателей эффективности классификации также использовалась площадь под ROC кривой AUC-ROC.

III. Этап 1. Выбор и определение сценариев предварительной обработки открытых табличных данных для выявления ССЗ у пациентов

Чтобы исследовать эффективность и точность модели AutoML, мы использовали пять наборов открытых данных из репозитория машинного обучения UCI (https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/), которые были объединены в таблицу по 12 общим показателям ССЗ и сформировали новую базу данных из 1190 наблюдений, доступную сегодня для исследовательских целей. Список показателей ССЗ пациентов, исключая метку класса заболевания (1: есть сердечные заболевания, 0: нет сердечных заболеваний), содержит: возраст, пол, максимальную частоту сердечных сокращений, давление в состоянии покоя, уровень холестерина, тип сердечной боли, гликемический индекс, признак стенокардии при физической нагрузке, признак ССЗ полученный по ЭКГ и два признака, характеризующих ST сегмент ЭКГ.

Анализ, очистка и преобразование данных на этапе предварительной обработки будут выполняться с использованием трех сценариев для ответа на второй исследовательский вопрос.

Первый сценарий S1 будет использовать встроенные возможности модели AutoML, которые включают автоматическое выполнение трудоемких ручных шагов: обработка недостающих данных, ручные преобразования объектов, разделение данных [10], а также удаление повторяющихся наблюдений и выбросов в данных, если таковые имеются.

Второй сценарий S2 включает предварительную обработку данных, реализованную отдельно от модели AutoML. В рамках этого сценария выполняются следующие шаги:

1) Статистический анализ агрегированных данных и визуализация распределений по каждому показателю ССЗ для проверки качества данных.
2) Очистка данных на основе выявления и удаления дубликатов данных, контекстуальных аномалий и выбросов, связанных с ошибками в значениях данных.
3) Преобразование данных в двоичные значения 0 и 1. Для этих целей значения непрерывных данных каждого показателя будут разделены на 5 интервалов, и каждый интервал будет представлен категориальным значением. Категориальные значения будут также преобразованы в значения 0 или 1. В результате количество признаков (столбцов в таблице) будет увеличено.
4) Деление набора данных на две части для обучения и тестирования AutoML и ML моделей в соотношении 80: 20.

Разница между вторым и третьим S3 сценариями заключается в способе преобразования очищенных данных. В третьем сценарии будет применено Z-нормализацию непрерывных значений к диапазону действительных чисел из интервала [0, 1], который рассчитывается как отношение разницы показателя и среднего значения в столбце, деленное на стандартное отклонение.

Статистический анализ данных, параметры которых были включены в используемый набор данных, показал, что средний возраст пациентов составил 53 года (минимальный возраст составлял 28 лет, а максимальный - 77 лет). Большинство пациентов были мужчинами, количество наблюдений с пометкой "имеет ССЗ" меньше, чем количество наблюдений с пометкой "не имеет ССЗ", визуальный анализ распределения данных подтвердил, что количество пациентов с ССЗ увеличивается с возрастом.

После устранения дубликатов окончательный набор данных составил 918 наблюдений. Визуальный анализ данных показал, что присутствует наличие выбросов по холестерину (19% от общего числа наблюдений). После удаления выбросов набор данных для обучения AutoML включал 714 очищенных наблюдений, которые были разделены в соотношении 80% для обучения и 20% для тестирования.

IV. Этап 2. Выбор модели AutoML и разработка с ее использованием системы классификации пациентов на тех, кто имеет и тех, кто не имеет ССЗ

В этой статье мы приняли решение использовать для построения модели AutoML фреймворк AutoGluon-Tabular [10], ориентированный на обработку табличных данных. Ключевые особенности AutoGluon-Tabular включают в себя устойчивость результатов классификации табличных данных гетерогенной природы, применение современных архитектур искусственной нейронной сети и автоматическую сборку ML моделей на основе многослойного стекинга и повторяющегося бэггинга. Всесторонняя эмпирическая оценка, приведенная авторами в статье [10] показывает, что фреймворк AutoGluon-Tabular значительно более точен, чем популярные фреймворки Auto ML, такие как Auto-WEKA [15], H2O [16], auto-sklearn [17]. Фреймворк AutoGluon-Tabular анализируя данные, создает

взвешенный ансамбль WeightedEnsemble_L2 с использованием 13 базовых ML моделей классификации данных [10]:

1) KNeighborsUnif и KNeighborsDist - это две модели k-ближайших соседей, которые отличаются способом определения соседей анализируемого экземпляра, среди которых и определяется его классификационная метка.
2) LightGBM, LightGBMXT, LightGBMLarge, XGBoost, CatBoost - это варианты моделей, для построения которых используется алгоритм градиентного бустинга, основанный на деревьях решений. Так, например, CatBoost создает модель прогнозирования в виде ансамбля деревьев решений, имеет механизм автоматической обработки пропущенных значений и поддержку параллельной обработки. XGBoost используется для работы с категориальными объектами. Он автоматически обрабатывает данные без необходимости предварительного кодирования, что упрощает процесс подготовки данных.
3) Ансамблевые модели типа случайного леса RandomForestGini, RandomForestEntr, ExtraTreesGini, ExtraTreesEntr включают несколько деревьев решений, для построения которых может быть использован индекс Джини или энтропия в данных как мера неопределенности, минимизация которой приводит к получению более чистых и информативных деревьев. Индекс Джини при построении деревьев используется чтобы уменьшить влияние выбросов и шума в данных.
4) Две модели глубокого обучения на основе искусственных нейронных сетей NeuralNetFastAI и NeuralNetTorch имеют высокоуровневые инструменты для обучения глубоких нейронных сетей. В статье [10] показано, что включение этих моделей позволяет улучшить точность классификации данных.

Генерируемая AutoML модель фреймворка AutoGluon-Tabular — это ансамблевая модель WeightedEnsemble_L2, второго уровня, которая объединяет взвешенные предсказания нескольких базовых ML моделей, имеющих наивысший рейтинг в результате обучения и оптимизации их гиперпараметров.

Модели ML первого уровня будут включены в модель второго уровня WeightedEnsemble_L2, если их точность на обучающей выборке равна или больше k, где k определяется алгоритмом AutoML. Такая модель будет обучаться с использованием k-кратной перекрестной валидации n раз для n различных случайных выборок входных данных, и эти параметры в AutoGluon-Tabular выбираются автоматически.

Для реализации системы классификации пациентов с целью выявления среди них страдающих ССЗ и проведения вычислительных экспериментов были использованы следующие технологии. Jupyter Notebook был применен для создания и выполнения кода, а также для документирования процесса анализа данных и построения модели. Google Colaboratory, онлайн-сервис, предоставляющий бесплатную облачную среду выполнения для блокнотов Jupyter с возможностью использования графических процессоров с высокой вычислительной нагрузкой (GPU) или тензорных процессоров (TPU). Для разработки и реализации всех этапов проекта software использовался язык программирования Python вместе с библиотеками:

1) Numpy - библиотека для обработки массивов данных.
2) Pandas - библиотека для работы с данными, включая загрузку, предварительную обработку и анализ.
3) Seaborn - библиотека для визуализации данных и создания информативных графиков.
4) Matplotlib.pyplot - библиотека для создания статических, анимированных и интерактивных визуализаций данных на Python.
5) Scikit-learn - библиотека машинного обучения, предоставляющая множество алгоритмов, моделей и показателей для построения и оценки моделей.
6) Autogluon - библиотека, предлагающая набор библиотек для выполнения автоматизированного машинного обучения, включая автоматическую подгонку модели и гиперпараметры.
7) Autogluon.tabular - библиотека, обеспечивающая автоматическое машинное обучение и выбор модели для задач анализа данных, представленных в виде таблиц.

V. Этап 3. Применение и сравнительная оценка моделей AutoML и базовых ML для выявления ССЗ на обучающей/тестовой выборках для различных сценариев предварительной обработки данных

Давайте рассмотрим результаты и точность классификации, полученные с помощью модели WeightedEnsemble_L2 для набора данных в трех сценариях предварительной обработки данных.

В первом сценарии модель WeightedEnsemble_L2 включала в себя две модели, NeuralNetTorch и LightGBMLarge.

S1: (встроенная в AutoML предобработка данных)
'stacker_info': {'num_base_models': 2, 'base_model_names': ['NeuralNetTorch', 'LightGBMLarge']} 'model_weights': {'NeuralNetTorch': 0.5, 'LightGBMLarge': 0.5}

Точность на тестовой выборке составила 88,81%, а на выборке для валидации модели - 87,8%. Выборка для валидации модели – это выборка данных из датасета без меток класса, которые сохранены отдельно. Они могут быть использованы для сравнения предсказания моделей. Результаты исследуемых моделей сценария S1, полученные на тестовых и валидационных данных, сведены в таблицу 1.

Во втором сценарии S2 для предварительной обработки данных модель WeightedEnsemble_L2 содержала только одну модель CatBoost, и точность модели AutoML улучшилась по сравнению со сценарием

S1 и составила 91,3% на валидациенной выборке и 92,3% на тестовой.

S2: (категоризация данных)
'stacker_info': {'num_base_models': 1, 'base_model_names': ['CatBoost']}, 'model_weights': {'CatBoost': 1.0}

Ниже приведена таблица 2, в которой показаны основные метрики модели WeightedEnsemble_L2 и моделей ML, которые она включила в свой взвешенный ансамбль в рамках сценария S2. В результате модели на основе бустинговых деревьев решений оказались лучшими на тестовой выборке, так как они

Таблица 1. Сравнение модели AutoML и моделей ML при использовании сценария S1 для предварительной обработки данных

№	модель	Stack level	метрика					
			accuracy (test)	accuracy (val)	roc_auc	precision	f1	recall
1	LightGBMXT	1	0,8881	0,8522	0,9377	0,8750	0,8974	0,9211
2	CatBoost	1	0,8811	0,8348	0,9405	0,8642	0,8917	0,9211
3	NeuralNetTorch	1	0,8811	0,8522	0,9305	0,8831	0,8889	0,8947
4	XGBoost	1	0,8811	0,8348	0,9379	0,8734	0,8903	0,9079
5	ExtraTreesEntr	1	0,8811	0,8261	0,9302	0,8642	0,8917	0,9211
6	LightGBM	1	0,8741	0,8522	0,9313	0,8718	0,8831	0,8947
7	ExtraTreesGini	1	0,8741	0,8435	0,9358	0,8625	0,8846	0,9079
8	LightGBMLarge	1	0,8671	0,8609	0,9195	0,8519	0,8790	0,9079
9	RandomForestGini	1	0,8671	0,8174	0,9272	0,8608	0,8774	0,8947
10	RandomForestEntr	1	0,8601	0,8174	0,9353	0,8500	0,8718	0,8947
11	NeuralNetFastAI	1	0,8252	0,8348	0,8879	0,8148	0,8408	0,8684
12	KNeighborsDist	1	0,6853	0,7043	0,7251	0,7123	0,6980	0,6842
13	KNeighborsUnif	1	0,6573	0,6957	0,7135	0,6901	0,6667	0,6447
14	WeightedEnsemble_L2	2	0,8881	0,8783	0,9307	0,8846	0,8961	0,9079

Таблица 2. Сравнение модели AutoML и моделей ML при использовании предварительной обработки набора данных в соответствии со сценарием S2

№	модель	Stack level	метрика					
			accuracy (test)	accuracy (val)	roc_auc	precision	f1	recall
1	ExtraTreesEntr	1	0,9650	0,8957	0,9863	0,9706	0,9635	0,9565
2	ExtraTreesGini	1	0,9580	0,8783	0,9863	0,9565	0,9565	0,9565
3	RandomForestEntr	1	0,9510	0,8783	0,9849	0,9559	0,9489	0,9420
4	RandomForestGini	1	0,9510	0,8870	0,9867	0,9559	0,9489	0,9420
5	CatBoost	1	0,9231	0,9130	0,9634	0,9265	0,9197	0,9130
6	LightGBMXT	1	0,8951	0,9043	0,9548	0,9219	0,8872	0,8551
7	NeuralNetTorch	1	0,8951	0,8957	0,9477	0,9219	0,8872	0,8551
8	LightGBM	1	0,8951	0,9043	0,9548	0,9219	0,8872	0,8551
9	XGBoost	1	0,8881	0,8783	0,9238	0,9077	0,8806	0,8551
10	LightGBMLarge	1	0,8601	0,8609	0,9555	0,9298	0,8413	0,7681
11	NeuralNetFastAI	1	0,8322	0,8609	0,9248	0,7922	0,8356	0,8841
12	WeightedEnsemble_L2	2	0,9231	0,9130	0,9634	0,9265	0,9197	0,9130

Таблица 3. Сравнение моделей AutoML и моделей ML при использовании предварительной обработки набора данных в соответствии со сценарием S3

№	модель	Stack level	метрика					
			accuracy (test)	accuracy (val)	roc_auc	precision	f1	recall
1	NeuralNetFastAI	1	0,8811	0,9478	0,9151	0,8308	0,8640	0,9000
2	LightGBMXT	1	0,8741	0,9478	0,9267	0,8281	0,8548	0,8833
3	ExtraTreesEntr	1	0,8741	0,9217	0,9241	0,8281	0,8548	0,8833
4	ExtraTreesGini	1	0,8671	0,9304	0,9271	0,8154	0,8480	0,8833
5	XGBoost	1	0,8531	0,9478	0,9086	0,8000	0,8320	0,8667
6	RandomForestEntr	1	0,8531	0,9304	0,9173	0,8000	0,8320	0,8667
7	RandomForestGini	1	0,8531	0,9217	0,9143	0,8095	0,8293	0,8500

8	CatBoost	1	0,8322	0,9391	0,9303	0,7903	0,8033	0,8167
9	NeuralNetTorch	1	0,8322	0,9304	0,9315	0,7903	0,8033	0,8167
10	LightGBMLarge	1	0,8252	0,9304	0,9072	0,7692	0,8000	0,8333
11	LightGBM	1	0,8182	0,9304	0,9096	0,7576	0,7937	0,8333
12	KNeighborsUnif	1	0,7133	0,7565	0,7599	0,6610	0,6555	0,6500
13	KNeighborsDist	1	0,7063	0,7565	0,7646	0,6552	0,6441	0,6333
14	WeightedEnsemble_L2	2	0,8741	0,9565	0,9259	0,8281	0,8548	0,8833

ориентированы на обработку категориальных данных.

Однако они значительно снизили свою точность на валидационной выборке, поэтому для получения устойчивого решения задачи классификации в итоговую модель WeightedEnsemble_L2 была включена более устойчивая модель CatBoost, относящаяся к моделям класса дерева решений, имеющая примерно одинаковые значения точности как на тестовой, так и на валидационной выборках данных.

Точность прогнозирования ССЗ с использованием модели WeightedEnsemble_L2 и ее базовых ML моделей, в которых использовался сценарий S3 для предварительной обработки данных, представлена в таблице 3.

Как следует из таблицы 3, обученная модель WeightedEnsemble_L2 модель на валидационном наборе показала точность 95,65%, а на тестовом наборе ее точность составила 87,41%. Две из 13 моделей, а именно LightGBMXT и ExtraTreesGini, были автоматически включены в ансамбль WeightedEnsemble_L2.

S3: (Z-нормализация)
'stacker_info': {'num_base_models': 2, 'base_model_names': ['LightGBMXT', 'ExtraTreesGini']} 'model_weights': {'LightGBMXT': 0.666, 'ExtraTreesGini': 0.333}

Проведенные исследования эффективности выявления ССЗ с использованием AutoML и ML моделей позволяют сформулировать ответы на вопросы исследования:

При сравнении AutoML и базовых ML моделей на валидационной выборке в первом и третьем сценариях точность модели AutoML превосходит точность базовых моделей, причем использование z-преобразования в сценарии S3, позволило получить максимальную точность в 95,65%. Во втором сценарии модель AutoML сравнима с точностью лучшей из оптимизированных моделей машинного обучения, которая в единственном числе и включена в состав WeightedEnsemble_L2. Это позволяет сделать вывод, что сценарий предобработки данных оказывает влияние на точность результатов классификации табличных данных, полученных в рамках модели AutoML при использовании фреймворка AutoGluon-Tabular, причем встроенный в эту библиотеку сценарий на исследуемом датасете (https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/), содержащем данные о показателях ССЗ пациентов, проигрывает сценариям предобработки данных, рассмотренным в этой статье.

I. Заключение

Автоматизированное машинное обучение позволяет упростить и ускорить процесс разработки приложений, использующих возможности моделей искусственного интеллекта, что является ключевым фактором в разработке ПОС. В данной статье были исследованы возможности AutoML фреймворка AutoGluon-Tabular [10] на примере модели WeightedEnsemble_L2 и 13 базовых ML моделей для прогнозирования ССЗ по 11 показателям здоровья пациентов с использованием датасета (https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/).

Результаты показали, что в зависимости от типа предварительной обработки данных алгоритм AutoML создает модели с разной структурой, которые реализуют прогнозы ССЗ с разной точностью. В нашем исследовании мы использовали три сценария предварительной обработки данных: (S1) встроенная автоматическая предварительная обработка данных, (S2) программно реализованная предварительная обработка данных с преобразованием значений в категориальные двоичные признаки и (S3) программно реализованная предварительная обработка данных с z-нормализацией непрерывных значений в диапазон [0,1]. Было показано, что точность модели WeightedEnsemble_L2, в которой использовалась встроенная предварительная обработка данных в AutoML, составила 87,8%. В то время как сгенерированная модель WeightedEnsemble_L2 на основе данных, подготовленных по второму сценарию, предсказала наличие ССЗ на обучающей выборке с точностью 91,3%. Модель AutoML, данные которой были предварительно обработаны согласно третьему сценарию, продемонстрировала точность 95,65%. Конечно, использование алгоритмов и библиотек для автоматического построения моделей машинного обучения имеет большое будущее и разнообразные применения в здравоохранении. Однако остаются вопросы, какую модель AutoML выбрать при решении конкретной клинической проблемы, как объяснить получаемое с ее помощью решение и по каким критерием можно доверять этим решениям.

Библиография

[1] Cai Y, Yu F, Kumar M, Gladney R, Mostafa J. Health Recommender Systems Development, Usage, and Evaluation from 2010 to 2022: A Scoping Review. Int J Environ Res Public Health. 2022 Nov 16;19(22):15115. doi: 10.3390/ijerph192215115. PMID: 36429832; PMCID: PMC9690602

[2] Chiang, P.H.; Wong, M.; Dey, S. Using Wearables and Machine Learning to Enable Personalized Lifestyle Recommendations to

Improve Blood Pressure. IEEE J. Transl. Eng. Health Med. 2021, 9, 2700513
[3] Gellert GA, Orzechowski PM, Price T, Kabat-Karabon A, Jaszczak J, Marcjasz N, Mlodawska A, Kwiecien AK, Kurkiewicz P. A multinational survey of patient utilization of and value conveyed through virtual symptom triage and healthcare referral. Front Public Health. 2023 Feb 2;10:1047291. doi: 10.3389/fpubh.2022.1047291. PMID: 36817183; PMCID: PMC9932322
[4] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab, Significance of machine learning in healthcare: Features, pillars and applications, International Journal of Intelligent Networks,Volume 3,2022,Pages 58-73, ISSN 2666-6030, https://doi.org/10.1016/j.ijin.2022.05.002
[5] Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms 2023, 16, 88. https://doi.org/10.3390/a16020088
[6] Javeed A, Khan SU, Ali L, Ali S, Imrana Y, Rahman A. Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions. Comput Math Methods Med. 2022 Feb 3;2022:9288452. doi: 10.1155/2022/9288452. PMID: 35154361; PMCID: PMC8831075
[7] Nashif, S., Raihan, Md. R., Islam, Md. R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 6, 854-873. https://doi.org/10.4236/wjet.2018.64057
[8] Seneviratne MG, Li RC, Schreier M, Lopez-Martinez D, Patel BS, Yakubovich A, Kemp JB, Loreaux E, Gamble P, El-Khoury K, Vardoulakis L, Wong D, Desai J, Chen JH, Morse KE, Downing NL, Finger LT, Chen MJ, Shah N. User-centred design for machine learning in health care: a case study from care management. BMJ Health Care Inform. 2022 Oct;29(1):e100656. doi: 10.1136/bmjhci-2022-100656. PMID: 36220304; PMCID: PMC9557254
[9] Rizzo G., & Lengelle R. The Evolution of Automated Machine Learning // IEEE Access, 2021, vol. 9, pp. 36595-36606
[10] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. 2020. ArXiv.org 2003.06505v1.
[11] Shashank Prasanna. Machine learning with AutoGluon, an open source AutoML library. 2020. AWS Open Source Blog. https://aws.amazon.com/ru/blogs/opensource/machine-learning-with-autogluon-an-open-source-automl-library/
[12] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020 Apr;104:101822. doi: 10.1016/j.artmed.2020.101822. Epub 2020 Feb 21. PMID: 32499001
[13] Sarra, Raniya & Dinar, Ahmed & Mohammed, Mazin. (2023). Enhanced accuracy for heart disease prediction using artificial neural network. Indonesian Journal of Electrical Engineering and Computer Science. 29. 375-383. 10.11591/ijeecs.v29.i1.pp375-383
[14] Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS ONE 14(5): e0213653. https://doi.org/10.1371/journal.pone.0213653
[15] Kotthoff,L.,Thornton,C.,Hoos,H.H.,Hutter,F.,and Leyton-Brown,K.Auto-WEKA2.0:Automaticmodel selectionandhyperparameteroptimizationinweka.JournalofMachineLearningResearch,18(25):1–5,2017.
[16] Pandey,P.ADeepDiveintoH2OsAutoML, 2019.URLhttp://www.h2o.ai/blog/a-deepdive-into-h2os-automl/.
[17] Feurer,M.,Klein,A.,Eggensperger,K.,Springenberg,J.T., Blum,M.,andHutter,F.Auto-sklearn:efficientandrobustautomatedmachinelearning.InAutomatedMachine Learning,pp.113–134.Springer,2019.

Афанасьева Татьяна Васильевна. Российский экономический университет имени Г.В. Плеханова, г. Москва, Россия. Доктор технических наук, доцент кафедры информатики. orcidID: 0000-0003-3779-7992, e-mail: tv.afanasjeva@gmail.com

Кузлякин Андрей Павлович. Российский экономический университет имени Г.В. Плеханова, г. Москва, Россия. Аспирант кафедры информатики. orcidID: 0009-0000-2721-9879, e-mail: andrey-kuzliakin@yandex.ru

Комолов Андрей Валерьевич. Российский экономический университет имени Г.В. Плеханова, г. Москва, Россия. Аспирант кафедры информатики. orcidID: 0009-0000-7196-2627, e-mail: komolov_1995@mail.ru

Comparative analysis of the accuracy of an automated machine learning model for detecting cardiovascular diseases

T.V. Afanasieva, A.P. Kuzlyakin, A.V. Komolov

***Abstract*—cardiovascular diseases (CVD) are widespread among patients with chronic non-communicable diseases and are one of the leading causes of mortality in the population, including those of working age. The development of patient-oriented systems for early detection of cardiovascular diseases using machine learning models is a promising direction that integrates medical knowledge and information intelligent technologies for medical decision support systems. To simplify and speed up the process of developing a specific solution based on a wide variety of machine learning models, the field of automatic machine learning (AutoML) is actively developing. The article provides a comparative analysis of the accuracy of an AutoML model created using the AutoGluon-Tabular library. The accuracy comparison was carried out in two directions: in relation to three scenarios for preprocessing data from patients with CVD and in relation to the basic machine learning models contained in the AutoGluon-Tabular library. A comparative analysis on the open UCI database showed that the accuracy of the AutoML model in identifying cardiovascular diseases varies from 87.41% to 95.65%, with the maximum accuracy obtained in the scenario with Z-normalization of the original data, and the minimum accuracy - when using data preprocessing algorithm built into AutoML.**

***Keywords*—cardiovascular diseases, automated machine learning, classification**

REFERENCES

[1] Cai Y, Yu F, Kumar M, Gladney R, Mostafa J. Health Recommender Systems Development, Usage, and Evaluation from 2010 to 2022: A Scoping Review. Int J Environ Res Public Health. 2022 Nov 16;19(22):15115. doi: 10.3390/ijerph192215115. PMID: 36429832; PMCID: PMC9690602

[2] Chiang, P.H.; Wong, M.; Dey, S. Using Wearables and Machine Learning to Enable Personalized Lifestyle Recommendations to Improve Blood Pressure. IEEE J. Transl. Eng. Health Med. 2021, 9, 2700513

[3] Gellert GA, Orzechowski PM, Price T, Kabat-Karabon A, Jaszczak J, Marcjasz N, Mlodawska A, Kwiecien AK, Kurkiewicz P. A multinational survey of patient utilization of and value conveyed through virtual symptom triage and healthcare referral. Front Public Health. 2023 Feb 2;10:1047291. doi: 10.3389/fpubh.2022.1047291. PMID: 36817183; PMCID: PMC9932322

[4] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab, Significance of machine learning in healthcare: Features, pillars and applications, International Journal of Intelligent Networks,Volume 3,2022,Pages 58-73, ISSN 2666-6030, https://doi.org/10.1016/j.ijin.2022.05.002

[5] Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms 2023, 16, 88. https://doi.org/10.3390/a16020088

[6] Javeed A, Khan SU, Ali L, Ali S, Imrana Y, Rahman A. Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions. Comput Math Methods Med. 2022 Feb 3;2022:9288452. doi: 10.1155/2022/9288452. PMID: 35154361; PMCID: PMC8831075

[7] Nashif, S., Raihan, Md. R., Islam, Md. R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 6, 854-873. https://doi.org/10.4236/wjet.2018.64057

[8] Seneviratne MG, Li RC, Schreier M, Lopez-Martinez D, Patel BS, Yakubovich A, Kemp JB, Loreaux E, Gamble P, El-Khoury K, Vardoulakis L, Wong D, Desai J, Chen JH, Morse KE, Downing NL, Finger LT, Chen MJ, Shah N. User-centred design for machine learning in health care: a case study from care management. BMJ Health Care Inform. 2022 Oct;29(1):e100656. doi: 10.1136/bmjhci-2022-100656. PMID: 36220304; PMCID: PMC9557254

[9] Rizzo G., & Lengelle R. The Evolution of Automated Machine Learning // IEEE Access, 2021, vol. 9, pp. 36595-36606

[10] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. 2020. ArXiv.org 2003.06505v1.

[11] Shashank Prasanna. Machine learning with AutoGluon, an open source AutoML library. 2020. AWS Open Source Blog. https://aws.amazon.com/ru/blogs/opensource/machine-learning-with-autogluon-an-open-source-automl-library/

[12] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020 Apr;104:101822. doi: 10.1016/j.artmed.2020.101822. Epub 2020 Feb 21. PMID: 32499001

[13] Sarra, Raniya & Dinar, Ahmed & Mohammed, Mazin. (2023). Enhanced accuracy for heart disease prediction using artificial neural network. Indonesian Journal of Electrical Engineering and Computer Science. 29. 375-383. 10.11591/ijeecs.v29.i1.pp375-383

[14] Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS ONE 14(5): e0213653. https://doi.org/10.1371/journal.pone.0213653

[15] Kotthoff,L.,Thornton,C.,Hoos,H.H.,Hutter,F.,and Leyton-Brown,K.Auto-WEKA2.0:Automaticmodel selectionandhyperparameteroptimizationinweka.JournalofMachineLearningResearch,18(25):1–5,2017.

[16] Pandey,P.ADeepDiveintoH2OsAutoML, 2019.URLhttp://www.h2o.ai/blog/a-deepdive-into-h2os-automl/.

[17] Feurer,M.,Klein,A.,Eggensperger,K.,Springenberg,J.T., Blum,M.,andHutter,F.Auto-sklearn:efficientandrobustautomatedmachinelearning.InAutomatedMachine Learning,pp.113–134.Springer,2019.

Afanasyeva Tatyana Vasilievna. Russian Economic University named after G.V. Plekhanov, Moscow, Russia. Doctor of Technical Sciences, Associate Professor of the Department of Computer Science. orcidID: 0000-0003-3779-7992, e-mail: tv.afanasjeva@gmail.com

Kuzlyakin Andrey Pavlovich. Russian Economic University named after G.V. Plekhanov, Moscow, Russia. Postgraduate student at the Department of Computer Science. orcidID: 0009-0000-2721-9879, e-mail: andrey-kuzliakin@yandex.ru

Komolov Andrey Valerievich. Russian Economic University named after G.V. Plekhanov, Moscow, Russia. Postgraduate student at the Department of Computer Science. orcidID: 0009-0000-7196-2627, e-mail: komolov_1995@mail.ru