Интеллектуальный анализ данных

AutoML: исследование существующих программных реализаций и определение общей внутренней структуры решений

И.А. Попова Г.И. Ревунков, Ю.Е. Гапанюк

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия

Аннотация. В статье рассматриваются различные программные реализации автоматизации процесса машинного обучения для решения задачи регрессии. Рассмотрено внутреннее устройство и возможности ряда существующих и обширно используемых инструментов автоматизированного машинного обучения таких, как LightAutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR. Возможности данных программных систем были исследованы для решения задачи регрессии на нескольких наборах данных. В результате исследования была выведена общая структура программного решения автоматизированного машинного обучения, которая может быть взята за основу при дальнейшем проектировании и разработке собственного программного продукта, а также проанализирована точность, с которой системы предсказывали значения целевого признака.
Ключевые слова: *автоматизированное машинное обучение, LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR, регрессия.*
DOI: 10.14357/20790279230106

Введение

AutoML является современной тенденцией в сфере машинного обучения. Данное направление активно исследуется научным сообществом, что подтверждается существованием и разработкой различных программных систем автоматизированного машинного обучения.

Специалисты в области машинного обучения регулярно сталкиваются с задачей выбора подходящего алгоритма с оптимальными гиперпараметрами для описания исследуемого набора данных. Для этого они обычно выполняют и оценивают множество конфигураций методом проб и ошибок. Однако для начинающих аналитиков данных это трудоемкая задача. Последние достижения в области исследований AutoML решают эту проблему путем автоматического поиска подходящего алгоритма с соответствующими гиперпараметрами. Основной задачей автоматизированного машинного обучения (AutoML) является автоматический поиск алгоритмов предварительной обработки входных данных и обучения выбранной системой модели с наилучшей производительностью обобщения на целевой (обрабатываемой) выборке.

Современные системы AutoML помогают автоматизировать практически весь процесс построения модели. На данный момент подлежат автоматизации следующие этапы моделирования: подготовка данных, обзор признаков, построение модели машинного обучения и оптимизация гиперпараметров, валидация построенной модели, построение отчетов, внедрение разработанной модели.

Алгоритмы машинного обучения для обработки и анализа данных комбинируются в конвейер, то есть применяются последовательно к обрабатываемой выборке. Далее происходит оптимизация этих алгоритмов путем подбора моделей и значений их гиперпараметров, причем большинство подходов выполняют оптимизацию всех параметров модели одновременно.

Однако стоит отметить, что системы автоматизированного машинного обучения широко применимы только для решения задач машинного обучения с учителем, то есть классификации и регрессии [1].

Каждая система отличается внутренней архитектурой, рядом возможностей, которые может использовать конечный пользователь и набором решаемых задач. Соответственно существующие решения по-разному могут справляться с одной и той же задачей машинного обучения.

На сегодняшний день направление AutoML развивается очень быстро, о чем свидетельствует большое количество публикаций. Примером соответствующих исследований являются [1, 3, 4, 7, 16, 17, 18]. Эти статьи не рассматривают производительность систем AutoML в отношении решения проблемы регрессии. Они не сравнивают окончательные модели, построенные системами на наборах данных, предназначенных для регрессии. Некоторые статьи [1, 3, 4, 7] описывают общие принципы концепции AutoML, но не учитывают конкретные системы AutoML либо авторы очень кратко приводят информацию в сравнительной таблице о некоторых системах AutoML. В работах [4, 7] даны очень краткие сведения о некоторых системах AutoML, которые более подробно рассмотрены в нашей статье (например, TPOT, H2O AutoML, Auto-Sklearn). В нашем исследовании даны рекомендации относительно применения рассмотренных систем AutoML при решении практических проблем. Вклад нашей работы заключается в том, что мы исследовали работу современных систем AutoML на примере задачи регрессии, описали подробно используемые системы, а также рассмотрели отечественное решение – LightAutoML в сравнении с существующими современными системами AutoML. Мы сравнили точность предсказания модели линейной регрессии с предсказаниями моделей, предложенных сложными системами AutoML. Чтобы оценить точность конечной модели, мы используем метрики MAE, RMSE, MedAE и R^2.

1. Цель исследования

В данной статье мы проведем сравнительный анализ ряда существующих программных систем автоматизированного машинного обучения: Light AutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR. Рассмотрим эффективность работы выбранных систем при решении задачи регрессии в нескольких предметных областях, используя ряд метрик точности (RMSE, MAE, MedAE, R^2).

2. Описание задачи

Основной задачей AutoML является автоматизация составления композиции моделей машинного обучения и дальнейшая параметризация подобранных алгоритмов, чтобы максимизировать значение выбранной метрики точности [4].

В данной статье сфокусируемся на использовании различных систем AutoML для решения задачи машинного обучения с учителем – линейной регрессии.

В процессе разработки модели машинного обучения с учителем необходимо предоставить на вход алгоритму набор признаков $\chi \subseteq \mathbb{R}^d$ и целевую переменную y. То есть модель обучается на образцах с известным значением целевой переменной y [3].

В сценариях машинного обучения используются подготовленные для последующего анализа и обработки статистические данные. Набор данных $D \subset \{(x, y) \mid x \in \chi, y \in Y\}$ является конечным отношением между пространством экземпляров и пространством меток, и мы обозначаем как D множество всех возможных наборов данных. Разработанная модель должна выдавать точные результаты для любых новых образцов данных. То есть, обобщающая способность является важнейшим свойством аналитической модели, приобретаемым в процессе обучения.

Зачастую процесс разработки модели машинного обучения представляет собой итеративный цикл обработки данных, обучения модели и ее оценки. Для того, чтобы получить на выходе удовлетворительную производительность модели, необходимо детально экспериментировать с различными комбинациями методов обработки данных, алгоритмов модели и гиперпараметров. Данный процесс достаточно затратен по времени и требует от исполнителя хороших знаний в области анализа данных.

Каждый этап приведенного на рис. 1 цикла разработки модели можно автоматизировать. При

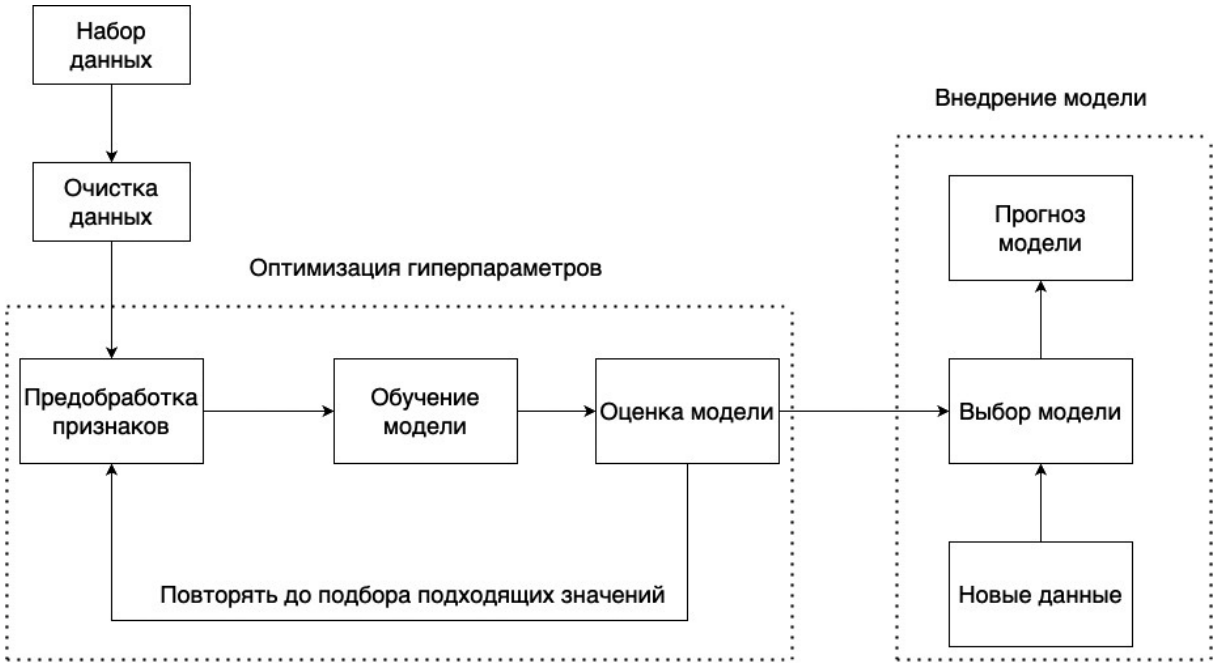


Рис. 1. Обобщенный цикл разработки модели машинного обучения

этом движение данных между модулями системы будет напоминать конвейер, состоящий из множества модулей: модуля обработки данных, выбора модели, построения отчетов, внедрения и т.д.

Основная цель машинного обучения с учителем состоит в том, чтобы найти конвейер, который минимизирует ошибку прогнозирования, усредненную по всем экземплярам выборки. Зачастую для оценки точности построенной модели используются такие метрики, как коэффициент детерминации (R^2), среднеквадратичная ошибка (RMSE), медианная абсолютная ошибка (MedAE), средняя абсолютная ошибка (MAE), AUC ROC, F1, log-loss и другие [7].

У каждой системы автоматизированного машинного обучения поиск и оптимизация алгоритмов осуществляется по-разному, о чем свидетельствуют различия во внутреннем устройстве систем. Исследование внутренней архитектуры систем автоматизированного машинного обучения позволит выявить общие закономерности в их построении, обнаружить места, которые могут быть усилены алгоритмами машинного обучения либо оптимизированы для обеспечения лучшей производительности решения.

3. Конвейер AutoML

Задачу AutoML можно сформулировать как процесс поиска f, который лучше обобщается в любом возможном T с минимальным вмешательством пользователя. Где f может быть композицией нескольких функций, которые могут преобразовывать входное пространство признаков, обрабатывать входные данные, строить модели и т. д. Такую композицию в контексте AutoML называют конвейером, который можно формализовать (формула 1).

$$f(x) = \upsilon_{\theta_\upsilon}(T_{\theta_T}(x)) \quad (1)$$

В данной формуле υ – модель машинного обучения, T – механизм преобразования признаков с гиперпараметрами θ_υ и θ_T соответственно [2]. Причем каждая из этих моделей может быть композицией других моделей.

Как показано на рис. 2 конвейер AutoML состоит из множества процессов обработки данных: предварительная обработка выборки, отбор признаков, генерация модели и ее оценка.

Первый этап включает в себя «добычу» информации, которая помогает повысить производительность сгенерированных моделей, создавая дополнительную информацию для изучения. На втором этапе к набору данных применяются различные методы работы с признаками: устранение пропусков в данных, кодирование категориальных признаков, масштабирование и нормализация признаков, обработка выбросов в данных. Далее подготовленные данные подаются на вход модулей, которые занимаются поиском наиболее эффективной архитектуры модели. Построение модели машинного обучения можно разделить на поиск пространства решений и оптимизацию гиперпараметров. Пространство поиска определяет принципы проектирования моде-

лей машинного обучения, которые можно разделить на две категории: традиционные модели (Linear Regression, SVM) и нейронные сети.

Методы оптимизации подразделяются на гиперпараметрическую оптимизацию (HPO) и оптимизацию архитектуры модели (AO). HPO оптимизирует параметры, связанные с обучением (например, скорость обучения и размер пакета), а AO оптимизирует параметры, связанные с моделью (например, количество слоев для нейронных архитектур и количество соседей для модели KNN) [5]. Завершающим этапом является тестирование производительности и выбор лучшей модели.

Зачастую первая часть конвейера не является последовательностью операций обработки данных, она имеет древовидную структуру с несколькими параллельными препроцессорами, которые затем объединяются. Оптимальный конвейер может быть реализован как строгий порядок, в котором должны применяться различные алгоритмы обработки данных, а также использоваться не более одного препроцессора каждого типа.

4. Обзор современных систем AutoML

Прогресс в области AutoML привел к появлению множества систем, которые автоматизируют проектирование и разработку моделей машинного обучения с учителем на разных этапах.

Рассмотрим ряд систем AutoML, которые используются в современных проектах машинного обучения.

4.1. LightAutoML

LightAutoML (LAMA) представляет собой решение с открытым исходным кодом, разработанное Sber AI Lab. Данная система позволяет автоматизи-

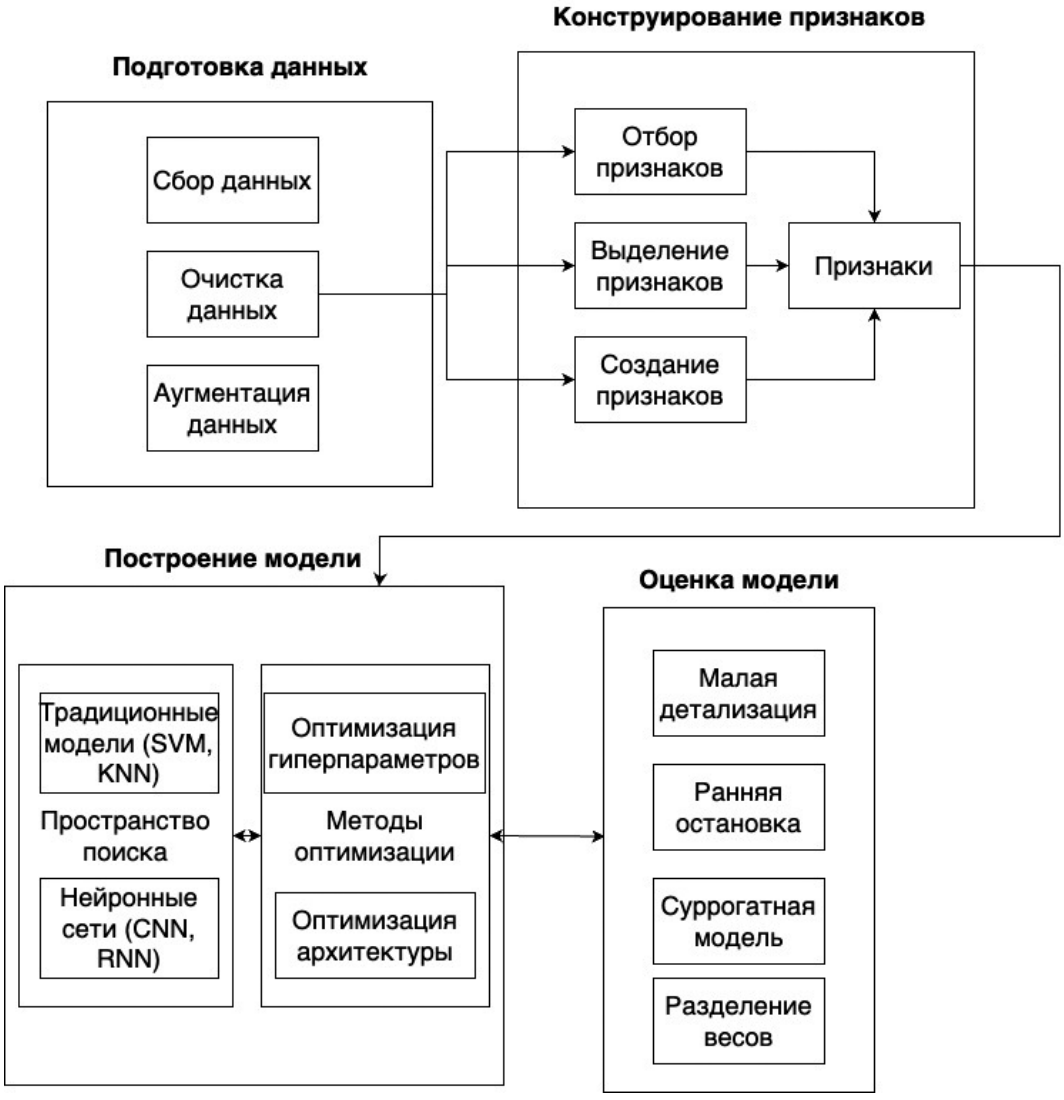


Рис. 2. Общая схема конвейера AutoML системы

ровать процесс построения модели для следующих задач: бинарная классификация, многоклассовая классификация, регрессия.

Разработанный конвейер дает возможность выполнять следующие операции: обрабатывать данные и автоматически настраивать гиперпараметры, строить отчеты, описывающие процесс разработки модели, конструировать собственные конвейеры из предоставляемых системой модулей, запускать модель в режиме предсказания.

Предлагаемая система LAMA работает только с двумя типами моделей - gradient boosted decision trees (GBMs) и линейными моделями, что значительно сокращает время без ущерба для производительности для решаемых типов задач и данных. LAMA состоит из модулей (пресетов), которые ориентированы на end-to-end разработку моделей для типичных задач ML. В настоящее время LightAutoML поддерживает следующие четыре предустановленных модуля:

1. TabularAutoML – фокусируется на классических задачах ML, работает с табличными наборами данных;
2. WhiteBox – решает задачу бинарной классификации с использованием простых интерпретируемых алгоритмов (логистическая регрессия);
3. NLP – способен комбинировать конвейер для обработки табличных данных с алгоритмами NLP (специальные средства извлечения признаков или предварительно обученные модели глубокого обучения);
4. CV – модуль для обработки изображений.

На рис. 3 представлен конвейер, который реализован в системе LightAutoML. Обязательными элементами конвейера являются:

- считыватель, который принимает на вход данные и выполняет их предварительную обработку, также на вход подается задача, которая подлежит решению с помощью системы;
- наборы данных, которые содержат метаданные и используются для валидации результатов; несколько конвейеров ML, которые складываются и / или усредняются с помощью Blender для получения единого прогноза.

Конвейеры ML могут быть вычислены независимо на одних и тех же наборах данных, а затем объединены вместе с использованием усреднения (или взвешенного усреднения).

4.2. TPOT

TPOT (Tree-based Pipeline Optimization Tool) является проектом на Python с открытым исходным кодом. TPOT автоматизирует конвейеры машинного обучения с помощью генетического программирования (GP), хорошо известного метода автоматического построения программ. В данной системе конвейер машинного обучения полностью автоматизирован и для определения оптимальной модели применяется генетический алгоритм.

Основное внимание в этом проекте уделяется обучению с учителем, а именно задаче классификации с поддержкой ста пятидесяти алгоритмов ScikitLearn [4], включая алгоритмы предварительной обработки. Система, как и Auto-Sklearn

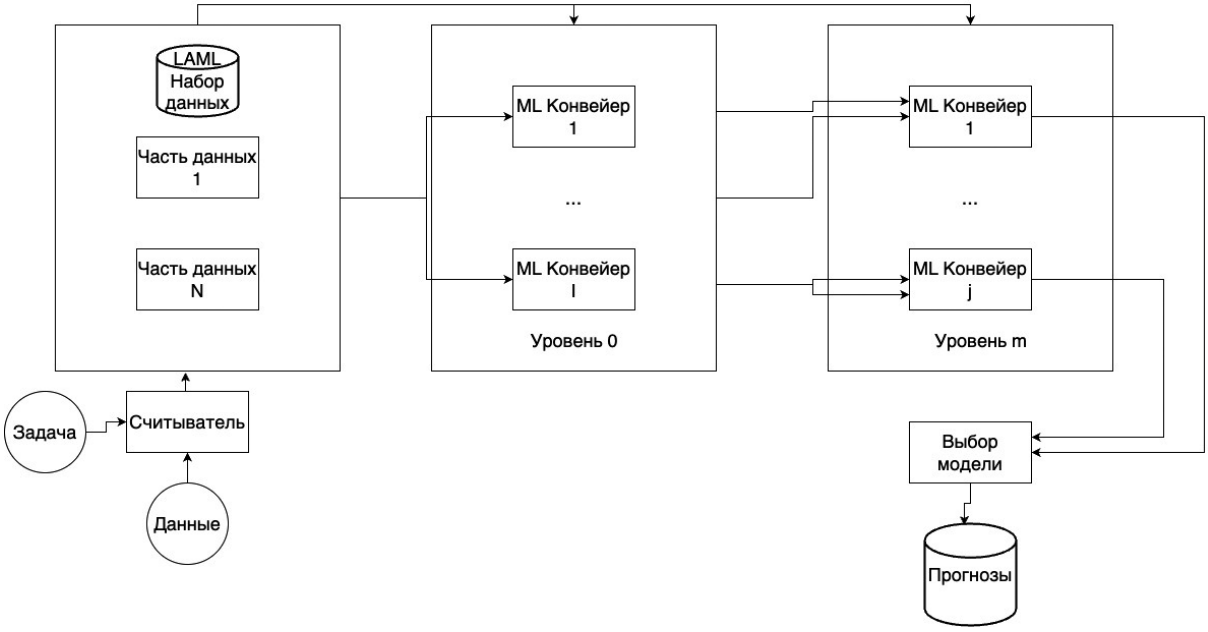


Рис. **3** Общая схема конвейера системы LAMA.

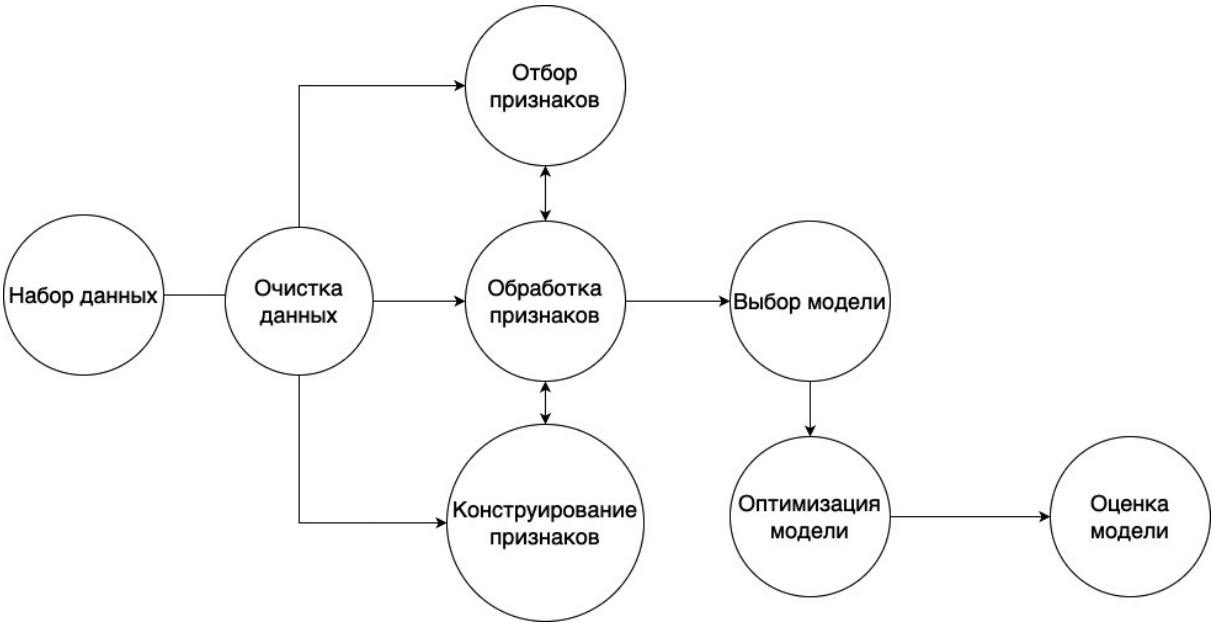


Рис. 4. Общая схема конвейера TPOT.

представляет собой надстройку над библиотекой Sklearn, однако использует собственные алгоритмы регрессии и классификации.

Отобраны двадцать лучших конвейеров с точки зрения максимальной точности перекрестной проверки и минимизации количества процессов, которые видоизменяются для создания конвейеров нового поколения. Каждый из конвейеров производит еще пять с помощью перекрестных методов или случайных вставок и усадок. Алгоритм выполняется для 100 поколений для каждого из обновляемых решений.

На рис. 4 представлен пример конвейера TPOT.

Система сохраняет копию набора данных и распараллеливает процесс обработки признаков и снижения размерности данных.

На данный момент TPOT не может работать с естественным языком и категориальными признаками.

4.3. Auto-Sklearn

Система написана на Python и использует алгоритмы и методы из программной библиотеки Scikit-Learn (15 алгоритмов классификации, 14 методов обработки признаков, 4 метода предварительной обработки данных) [10]. Auto-Sklearn реализует алгоритм SMAC для настройки гиперпараметров. Система умеет генерировать признаки, выбирать модель, настраивать гиперпараметры. Auto-Sklearn представляет два новшества: использование метаобучения для запуска процедуры байесовской оптимизации и шаг построения ансамбля, который использует более одной конфигурации, найденной в процедуре оптимизации.

Auto-Sklearn запускает процедуру байесовской оптимизации, предоставляя начальные экземпляры из конфигураций, которые дали наилучшие результаты для аналогичных наборов данных [6].

На рис. 5 представлен пример конвейера Auto-Sklearn. Настройка гиперпараметров и предварительная обработка данных частично выполняются с помощью метаобучения. Результаты метаобучения предопределяют пространство поиска. Конвейер AutoML использует это пространство поиска итеративно, начиная с процессора данных, за которым следует препроцессор обработки признаков, за которым следует обучение классификатора либо регрессора. Результаты оцениваются, и гиперпараметры оптимизируются с помощью байесовского оптимизатора. Зачастую лучшие значения производительности показывают ансамблевые модели.

4.4. H2O AutoML

H2O является системой машинного обучения с открытым исходным кодом, с распределенной памятью и линейной масштабируемостью. Текущая версия системы AutoML может обучать и выполнять кросс-валидацию для случайного леса, градиентного бустинга, глубоких нейронных сетей, а затем обучать составной ансамбль, используя все модели. H2O AutoML автоматизирует процесс построения большого количества моделей, чтобы выявить наиболее эффективную модель.

Ключевые особенности H2O AutoML: среда с открытым исходным кодом, которая предоставляет распределенные реализации методов машинного обучения; система реализована на Java, однако есть API для таких языков, как Python, R, Scala, а также

Рис. 5. Общая схема конвейера Auto-Sklearn

доступен веб-интерфейс; H2O AutoML работает на таких платформах, как Hadoop, Spark, AWS.

4.5. MLJAR

Систему AutoML MLJAR можно использовать для создания полного конвейера машинного обучения с конструированием признаков и настройкой гиперпараметров. MLJAR поддерживает следующие алгоритмы машинного обучения: нейронные сети, XGBoost, Catboost, LightGBM и другие. MLJAR строит несколько моделей в зависимости от выбранных алгоритмов и рассчитывает окончательный прогноз путем объединения в ансамбль или стекинга моделей.

На рис. 6 представлен пример конвейера MLJAR.

4.6. Сравнительный анализ AutoML систем

Системы AutoML различаются по своему внутреннему устройству и, соответственно, функциям, которые они предоставляют пользователю. Выделим отличительные особенности рассмотренных систем и дадим краткие рекомендации по их применению:

- Система LightAutoML эффективна для решения таких задач, как бинарная или многоклассовая классификация, а также регрессия, где входные данные могут содержать одновременно различные типы признаков: числа, тексты, категориальные данные, даты. Можно применять LightAutoML в качестве инструмента быстрой проверки гипотез, а также построения моделей машинного обучения, которые будут описаны с помощью линейных моделей и деревьев решений с градиентным бустингом. Также возможно расширять систему собственными модулями обработки данных, таким образом настраивая LightAutoML под решение новых задач.
- Система TPOT не может обрабатывать пропуски в наборе данных, а также не может работать с нечисловыми признаками. Пользователь должен предварительно обработать данные, только потом подавать их в систему. На предварительно обработанных данных TPOT позволяет строить модели классификации или регрессии, используя алгоритмы Sklearn. Поэтому данная система подойдет вам, если вы работаете с табличными данными, не содержащими пропущенные значения, а также все признаки объектов выборки являются числовыми. Иначе требуется предварительная экспертиза и обработка данных.
- Система H2O AutoML предоставляет метод импорта файлов, который позволяет загружать табличные данные, состоящие из категориальных и числовых признаков, а затем, используя внутреннюю эвристику, делит данные на подвыборки для обучения и тестирования целевой модели. Данная система автоматизирует такие этапы, как предварительная обработка данных, обучение и настройка модели, объединение различных моделей, чтобы выбрать модели с наилучшей производительностью (зачастую данные описываются ансамблями GBM, GLM, DNN моделей). H2O предоставляет удобный

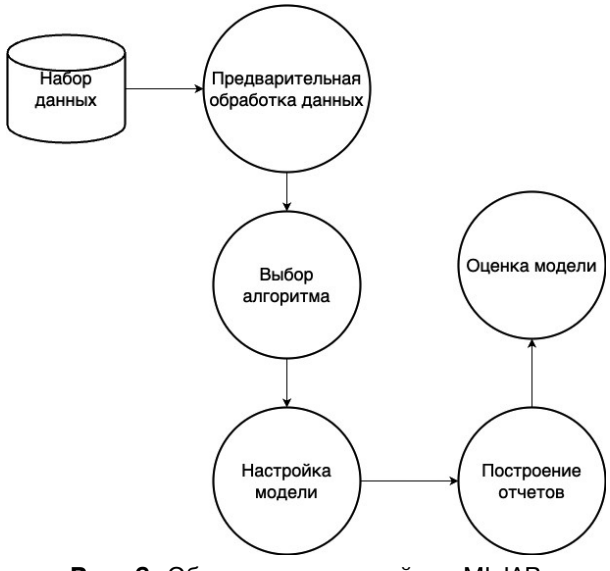


Рис. 6. Общая схема конвейера MLJAR

пользовательский интерфейс H2O Flow – интерактивную веб-среду, которая позволяет совмещать выполнение кода, математические вычисления, графики и мультимедиа в одном документе. Использовать данную систему рекомендуется, если выборка содержит пропуски, категориальные и числовые признаки, также если необходимо проанализировать предложенные системой модели, представленные в виде таблицы с метриками качества (RMSE, MSE, MAE и др.).

- Система MLJAR довольно быстро справляется с построением и обучением целевой модели на выборках разного размера. Она может создавать отчеты с помощью разметки Markdown, содержащие сведения о процессе обучения моделей, работая со множеством разных моделей машинного обучения, обрабатывать пропуски в данных и работать с большим количеством типов признаков. Если вам нужно быстро построить модель машинного обучения на данных, содержащих пропуски и большое количество признаков различных типов, то вы можете выбрать MLJAR.
- Auto-Sklearn не может обрабатывать пропущенные значения, но быстро справляется с задачей выбора самой эффективной модели машинного обучения. Эта система построена поверх алгоритмов машинного обучения из библиотеки Sklearn. Auto-Sklearn сочетает в себе методы, которые помогут создать модель с настроенными гиперпараметрами, но пользователю придется предварительно предобработать входные данные. Auto-Sklearn использует байесовские методы оптимизации для поиска наиболее производительного конвейера для заданного набора данных, поэтому вычисления даже на больших наборах данных будут производиться достаточно быстро.

5. Исследование систем AutoML для задачи регрессии

Мы сравнили рассмотренные выше системы AutoML, решая задачу регрессии на нескольких наборах данных, и оценили при помощи ряда метрик точности конечные результаты работы систем.

Для экспериментальной части исследования были выбраны два набора данных:

1) cars – выборка, содержащая 301 объект, в качестве целевого признака выступает цена автомобиля;
2) powerplant - набор данных, содержащий параметры 9568 предприятий, где в качестве целевого признака выступает почасовая выработка электроэнергии (МВт/ч).

Набор данных cars опубликован на ресурсе Kaggle [11], а набор данных plants опубликован в репозитории машинного обучения UCI [12].

Исходные данные были обработаны 5 системами автоматизированного обучения: LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR.

В итоге данными системами был построен конвейер для обработки данных, определены статистические модели и настроены их гиперпараметры. Затем нами была произведена оценка точности каждой итоговой модели.

Опишем метрики, которые были использованы для оценки точности конечной модели, построенной выбранными системами AutoML:

1. Средняя абсолютная ошибка (MAE). Эта метрика не чувствительна к выбросам в наборе данных, но она не нормирована.
2. Среднеквадратическая ошибка (RMSE). Данная метрика чувствительна к выбросам.
3. Средняя абсолютная ошибка (MedAE). Данная метрика не чувствительна к выбросам в наборе данных.
4. Коэффициент детерминации (R^2). Выбросы существенно влияют на коэффициент детерминации.

Соответственно, чем ближе значение метрик к нулю, тем точнее работает модель (кроме метрики R^2, которая принимает значения в диапазоне от 0 до 1, чем ближе значение R^2 к единице, тем точнее модель).

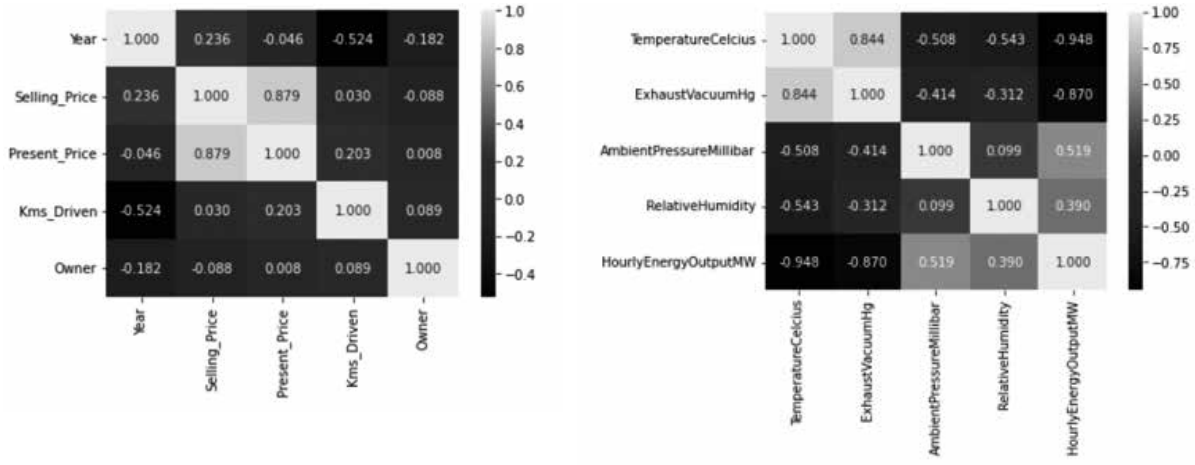
В работе используются библиотечные реализации описанных выше метрик, которые предоставляет модуль metrics пакета sklearn.

5.1. Анализ выборок

Для того, чтобы предварительно оценить внутреннюю организацию данных и построить гипотезы, визуализируем экспериментальные выборки.

Рис. 7а и 7б отображают корреляционную матрицу, которая содержит коэффициенты корреляции между парами признаков из набора данных. Матрица корреляции содержит на главной диагонали единицы, то есть является симметричной. В данном случае при расчете матрицы корреляции использовался коэффициент корреляции Пирсона. Для удобного восприятия корреляционной матрицы используем «тепловую карту», где при помощи цветов окрашены ячейки, содержащие коэффициенты корреляции.

Для набора данных cars на целевой признак «Selling Price» больше всего влияет признак «Present Price» (коэффициент корреляции равен 0,844). В наборе данных plants признак «ExhaustVacuumHg» коррелирует с целевым признаком «HourlyEnergyOutputMW» с коэффициен-



а б

Рис. 7. Корреляционные карты

том -0,87. В данном случае важен модуль коэффициента корреляции, который показывает силу зависимости между признаками, отрицательный знак указывает на обратную связь между признаками, что для рассматриваемой задачи не имеет значения.

Проанализировав матрицу корреляции, можно сделать вывод, что данные в выбранных датасетах можно аппроксимировать прямой линией, следовательно, можно предположить, что модели линейной регрессии будет достаточно, чтобы добиться высокой точности предсказаний.

5.2. Анализ полученных результатов

В результате эксперимента, с помощью описанных выше AutoML систем, были определены модели машинного обучения, наилучшим образом описывающие входные данные. При этом каждой из систем AutoML был построен конвейер, который включал в себя все этапы обработки данных, подбора модели и настройки ее гиперпараметров.

На рис. 8-9 представлены гистограммы, на которых можно увидеть значения погрешности предсказания целевого признака для каждой исследуемой в работе AutoML системы. На рисунке 8-9 представлены значения метрик качества для тестовой выборки набора данных cars. Можно заметить, что наиболее точной оказалась модель линейной регрессии по оценкам MAE, RMSE, MedAE, также хороших результатов позволила добиться модель, выбранная системой MLJAR (метрики MAE, RMSE, MedAE).

На рис. 8 показаны показатели качества для тестовой выборки набора данных об автомобилях. Видно, что модель линейной регрессии по оценкам MAE, RMSE и MedAE оказалась наиболее точной,

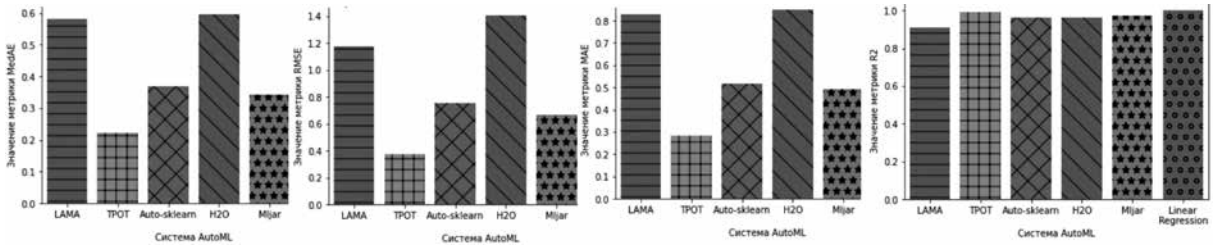


Рис. 8. Значения метрик для набора данных cars

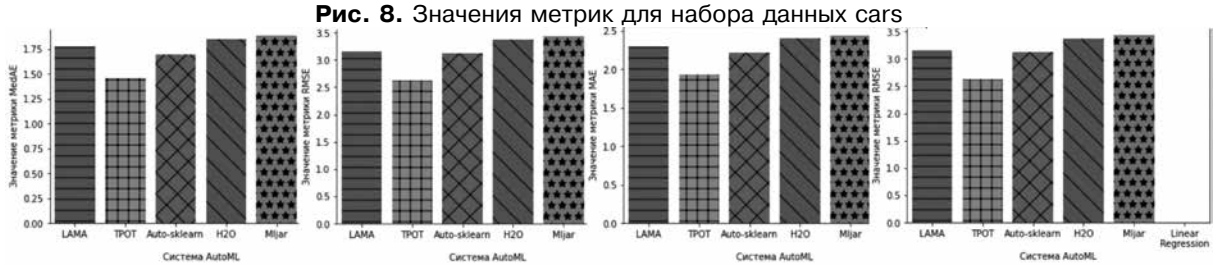


Рис. 9. Значения метрик для набора данных plants

а модель, выбранная по системе MLJAR (метрики MAE, RMSE, MedAE), также показала хорошие результаты. Поскольку мы используем 4 метрики для оценки точности окончательной модели машинного обучения, лучшая система AutoML определяется с помощью анализа Парето.

На рис. 9 показаны оценки качества для тестовой выборки набора данных об автомобилях. Видно, что модель линейной регрессии по оценкам MAE и RMSE оказалась наиболее точной, а модель, выбранная по системе TPOT (метрики MAE, RMSE и R^2), также показала хорошие результаты.

Выбранные системы показывают совершенно разные результаты при решении задачи регрессии на выборках разного размера. Набор данных автомобилей содержит 301 объект, а набор данных растений содержит 9568 объектов. Наилучшие результаты показывают системы TPOT и MLJAR (если учитывать значения метрик MAE, RMSE, MedAE).

Система LAMA неплохо справляется с задачей подбора и настройки модели, однако автоматизация всего процесса занимает длительное время (в среднем до 5 минут). Это связано с большим количеством различных расчетов, производимых системой из-за обилия предоставляемых ею возможностей.

В табл. 1 показано время, затрачиваемое каждой системой AutoML на выбор и обучение окончательной модели машинного обучения.

Табл. 1

Время работы систем AutoML

Название	Набор данных cars	Набор данных plants
Auto-Sklearn	2 мин 1 сек	1 мин 59 сек
TPOT	31.2 сек	1 мин 30 сек
MLJAR	29.7 сек	54.4 сек
H2O	1 мин 2 сек	1 мин 2 сек
LAMA	3 мин 18 сек	5 мин 40 сек
Linear Regression	19.8 мс	6.78 сек

На основании анализа Табл. 1 можно сделать вывод, что системе LightAutoML потребовалось больше всего времени для построения оптимального конвейера, выбора наилучшей модели и ее обучения. Модель линейной регрессии оказалась самой быстрой для построения, а система MLJAR выполняла вычисления быстрее, чем другие системы.

Заключение

Таким образом, в рамках данной работы формализован процесс оптимизации построения конвейеров данных и настройки алгоритмов машинного обучения. Рассмотрен ряд автоматизированных систем машинного обучения: Light AutoML (LAMA), TPOT, Auto-Sklearn, MLJAR, H2O AutoML. Исследован процесс построения модели регрессии с использованием перечисленных систем для нескольких наборов данных, содержащих категориальные и числовые признаки.

Литература

1. *Nagarajah, T., Poravi, G.* A Review on Automated Machine Learning (AutoML) Systems. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–6. Pune, India (2019). https://doi.org/10.1109/I2CT45611.2019.9033810
2. *Bahri, M., Salutari, F., Putina, A. et al.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. Int J Data Sci Anal (2022). https://doi.org/10.1007/s41060-022-00309-0
3. *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities. ACM Computing Surveys (CSUR) 54, 1–36 (2022)
4. *He X., Zhao K., Chu X.* AutoML: A survey of the state-of-the-art. Knowl. Based Syst., 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622
5. *Escalante, H.J.* Automated Machine Learning – a brief review at the end of the early years. arXiv:2008.08516. https://doi.org/10.48550/arXiv.2008.08516
6. *Bahri, M., Salutari, F., Putina, A., Sozio, M.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics, Springer Verlag, 2022. https://doi.org/10.1007/s41060-022-00309-0
7. *Koroteev, M.V.* Review of some modern trends in machine learning technology. E-Management 1(1), 26–35 (2018)
8. *Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M.T., Hutter, F.* Practical Automated Machine Learning for the AutoML Challenge 2018. In: International Workshop on Automatic Machine Learning at ICML, pp. 1189-1232 (2018)
9. Car Dekho Data, https://www.kaggle.com/datasets/shindenikhil/car-dekho-data. Last accessed 12 December 2022
10. Combined Cycle Power Plant Dataset, https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant. Last accessed 12 December 2022
11. LightAutoML – Automatic model creation framework, https://github.com/sb-ailab/

LightAutoML. Last accessed 12 December 2022
12. TPOT – A Python Automated Machine Learning tool, https://github.com/EpistasisLab/tpot. Last accessed 12 December 2022
13. Auto-Sklearn – An automated machine learning toolkit, https://github.com/automl/auto-sklearn. Last accessed 12 December 2022
14. H2O AutoML – Open-Source Automated Machine Learning, https://h2o.ai/platform/h2o-automl/. Last accessed 12 December 2022
15. MLJAR – Automate your Machine Learning pipeline, https://mljar.com/. Last accessed 12 December 2022
16. Chen, Yi-Wei, Qingquan Song, and Xia Hu.: Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 35-50 (2021).
17. Elshawi, Radwa, Mohamed Maher, and Sherif Sakr: Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287 (2019).
18. V*akhrushev, Anton, et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. arXiv preprint arXiv:2109.01528 (2021).

Попова Инна Андреевна. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Аспирант. Количество печатных работ: 5. Область научных интересов: информационные технологии, машинное обучение. E-mail: popovai1@student.bmstu.ru

Гапанюк Юрий Евгеньевич. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Доцент. Количество печатных работ: более 100. Область научных интересов: проектирование автоматизированных систем, проектирование гибридных интеллектуальных информационных систем, сложные графовые модели. E-mail: gapyu@bmstu.ru (ответственный за переписку)

Ревунков Георгий Иванович. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Доцент. Количество печатных работ: более 100 (3 монографии). Область научных интересов: проектирование автоматизированных систем, проектирование гибридных интеллектуальных информационных систем, базы данных. E-mail: revunkov@bmstu.ru

AutoML: Examining Existing Software Implementations and Determining the Overall Internal Structure of Solutions

I.A. Popova , G.I. Revunkov, Y.E. Gapanyuk

Bauman Moscow State Technical University, Moscow, Russia

Abstract. The article discusses various software implementations of the process of automating the task of using machine learning to solve the linear regression problem. The internal structure and capabilities of a number of existing and widely used automated machine learning tools such as LightAutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR are considered. The capabilities of these software systems have been explored to solve the regression problem on multiple datasets.

Keywords: *automated machine learning (AutoML), LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR, Regression.*

DOI: 10.14357/20790279230106

References

1. *Nagarajah, T., Poravi, G.* A Review on Automated Machine Learning (AutoML) Systems. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–6. Pune, India (2019). https://doi.org/10.1109/I2CT45611.2019.9033810
2. *Bahri, M., Salutari, F., Putina, A. et al.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. Int J Data Sci Anal (2022). https://doi.org/10.1007/s41060-022-00309-0

3. *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities. ACM Computing Surveys (CSUR) 54, 1–36 (2022)
4. *He X., Zhao K., Chu X.* AutoML: A survey of the state-of-the-art. Knowl. Based Syst., 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622
5. *Escalante, H.J.* Automated Machine Learning – a brief review at the end of the early years. arXiv:2008.08516. https://doi.org/10.48550/arXiv.2008.08516
6. *Bahri, M., Salutari, F., Putina, A., Sozio, M.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics, Springer Verlag, 2022. https://doi.org/10.1007/s41060-022-00309-0
7. *Koroteev, M.V.* Review of some modern trends in machine learning technology. E-Management 1(1), 26–35 (2018)
8. *Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M.T., Hutter, F.* Practical Automated Machine Learning for the AutoML Challenge 2018. In: International Workshop on Automatic Machine Learning at ICML, pp. 1189-1232 (2018)
9. Car Dekho Data, https://www.kaggle.com/datasets/shindenikhil/car-dekho-data. Last accessed 12 December 2022
10. Combined Cycle Power Plant Dataset, https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant. Last accessed 12 December 2022
11. LightAutoML – Automatic model creation framework, https://github.com/sb-ailab/LightAutoML. Last accessed 12 December 2022
12. TPOT – A Python Automated Machine Learning tool, https://github.com/EpistasisLab/tpot. Last accessed 12 December 2022
13. Auto-Sklearn – An automated machine learning toolkit, https://github.com/automl/auto-sklearn. Last accessed 12 December 2022
14. H2O AutoML – Open-Source Automated Machine Learning, https://h2o.ai/platform/h2o-automl/. Last accessed 12 December 2022
15. MLJAR – Automate your Machine Learning pipeline, https://mljar.com/. Last accessed 12 December 2022
16. Chen, Yi-Wei, Qingquan Song, and Xia Hu.: Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 35-50 (2021).
17. Elshawi, Radwa, Mohamed Maher, and Sherif Sakr: Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287 (2019).
18. V*akhrushev, Anton, et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. arXiv preprint arXiv:2109.01528 (2021).

Popova Inna Andreevna. Graduate student, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: 5. Research interests: information technology, machine learning. E-mail: popovai1@student.bmstu.ru

Gapanyuk Yuriy Evgenievich. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100. Research interests: designing of automated systems, designing of hybrid intelligent information systems, complex graph models. E-mail: gapyu@bmstu.ru

Revunkov Georgiy Ivanovich. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100 (3 monographs). Research interests: designing of automated systems, designing of hybrid intelligent information systems, database systems. E-mail: revunkov@bmstu.ru