СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ
SYSTEM ANALYSIS AND PROCESSING OF KNOWLEDGE

УДК 004.85
DOI 10.52575/2687-0932-2021-48-2-350-359

Метод выбора модели машинного обучения на основе устойчивости предикторов с применением значения Шепли

Воробьев А.В.
Курский государственный университет
Россия, г. Курск, ул. Радищева, 33
E-mail: 505216@inbox.ru

Аннотация. В статье рассмотрено использование вектора Шепли в регрессионном анализе как метода, снижающего дестабилизирующие воздействие мультиколлинеарности предикторов, а также его применение в интерпретации машинного обучения. Определены ограничения его применения. На основе значений Шепли предложен метод выбора стабильной модели, позволяющий стабилизировать показатели точности моделей при искажении предикторов и шумов, повышать показатели точности при снижении набора данных на классических и модернизированных ансамблевых алгоритмах. Испытания разработанного алгоритма проводились как на синтезированных, так и на общедоступных популярных DataSet для машинного обучения, с различной размерностью по количеству атрибутов и периодам наблюдений. В экспериментах наблюдался устойчивый положительный эффект, выраженный в сокращении взвешенной абсолютной процентной ошибки прогнозирования и рост данного эффекта при увеличении атрибутной размерности выборки. Предложенный алгоритм может выступать в качестве инструмента повышения эффективности ансамблевых алгоритмов машинного обучения, в том числе в высокоэффективных и высокоскоростных.

Ключевые слова: машинное обучение, ансамблевые алгоритмы, значение Шепли, точность модели, устойчивость к шуму данных.

Для цитирования: Воробьев А.В. 2021. Метод выбора модели машинного обучения на основе устойчивости предикторов с применением значения Шепли. Экономика. Информатика, 48 (2): 350–359. DOI 10.52575/2687-0932-2021-48-2-350-359.

Feautre stability based machine learning model selection method with usage of Shapley values

Alexander V. Vorobyev
Kursk State University
Russia, Kursk, 33 Radisheva St,
E-mail: 505216@inbox.ru

Abstract. In this article the usage of Shapley vectors in regression analysis as a method to reduce destabilizing impact of feature multicollinearity and its usage in interpreting machine learning results is considered. The limitations of its application were defined. A Shapley value based method of stable model selection, allowing for stabilization of models' precision in event of feature and noise distortion, and for increasing precision of classic and innovative ensemble algorithms while shrinking the dataset is proposed. The developed algorithm was tested on both synthetic and publicly available popular machine learning datasets with different amounts of attributes and observation periods. The experiments showed a stable positive effect of decreasing WMAPE and increasing of the effect upon increasing the feature amount of sampling. The suggested algorithm can be

used as a tool to increase the efficiency of the ensemble machine learning algorithms, including the high-speed and high-efficiency algorithms.

Keywords: machine learning, ensemble algorithms, Shapley value, model precision, data noise resistance.

For citation: Vorobev A. 2021. Feautre stability based machine learning model selection method with usage of Shapley values. Economics. Information technologies, 48 (2): 350–359. (in Russian). DOI 10.52575/2687-0932-2021-48-2-350-359.

Введение

Увеличение доступности и широкое внедрение машинного обучения как эффективного инструмента прогнозирования определяют особое внимание научного сообщества к проблемам выбора предикторов. Наряду с требованиями к высокой точности моделей формируется потребность в устойчивости к шуму в данных.

В данной статье мы предлагаем и анализируем новую и эффективную стабильную процедуру выбора предикторов, обеспечивающую рост точности и повышения устойчивости моделей машинного обучения, основанных на ансамблевых алгоритмах.

Мультиколлинеарность предикторов и значение Шепли

Одной из основных причин нестабильности прогностических моделей является мультиколлинеарность предикторов [Mason, 1991]. В последнее время одной из наиболее признанных концепций решения представленной проблемы в литературе является значение Шепли [Landinez-Lamadrid и др., 2017]. Применение данного метода в регрессионном анализе позволяет получить коэффициенты скорректированной регрессии, не подверженные дестабилизирующим воздействиям линейной зависимости между предикторами и влиянию случайных величин [Михеенко и др., 2020].

Подход Шепли состоит в том, чтобы рассмотреть пространство всех игр, в которые может играть некоторый потенциально очень большой набор игроков (U). В конкретной игре v фактически вовлеченные игроки содержатся в любой выборке, которая является подмножеством N из числа U, такой, что $v(S) = v(S \cap N)$ для любого подмножества игроков $S \subset U$. Если выборка N для игры v не содержит некоторого игрока i, то i является нулевым игроком, потому что i не влияет на выигрыш $v(S)$ любой коалиции S. Таким образом, любой набор, содержащий выборку, сам по себе является выборкой игры, и всякий игрок, не входящий во все выборки, является нулевым игроком.

Шепли определил *значение* для игр как функцию, которая присваивает каждой игре v число $\phi_i(v)$ для каждого i в U. Он предложил, чтобы такая функция подчинялась трем аксиомам. Аксиома *симметрии* требует, чтобы позиции игроков не имели роли в определении значения, которое должно быть чувствительным только к тому, как характеристическая функция реагирует на присутствие игрока в коалиции. В частности, аксиома симметрии требует, чтобы игроки, к которым одинаково относится характеристическая функция, одинаково относились и к значению.

Вторая аксиома, обычно называемая аксиомой *выборки*, определяет, чтобы сумма $\phi_i(v)$ по всем игрокам i в любом векторе N равнялась $v(N)$. Поскольку это должно иметь место для любого вектора, это означает, что $\phi_i(v) = 0$, если i является нулевым игроком в v. Иногда эта аксиома определяется как состоящая из двух частей: аксиомы *эффективности* ($\sum_{i \in N} \phi_i(v) = v(N)$ *для некоторого вектора* N) и аксиомы *нулевого игрока* («Аксиомы болвана»).

Третья аксиома – *аддитивности*, требует, чтобы для любых игр v и w, $\phi(v) + \phi(w) = \phi(v + w)$ *(то есть* $\phi_i(v) + \phi_i(w) = \phi_i(v + w)$ для всех i в U, где игра $[v + w]$ определяется как $[v + w](S) = v(S) + w(S)$ для любой коалиции S). Эта аксиома определяющая, как значения различных игр должны быть связаны друг с другом, является демонстрацией существования

уникальной функции ϕ, определенной на пространстве всех игр, которая удовлетворяет этим трем аксиомам.

Самый простой способ понять, почему эта функция существует и уникальна, – это представить характеристическую функцию *v* как вектор с 2^U-1 компонентами, по одному для каждого непустого подмножества U. Тогда множество *G* всех (не обязательно супераддитивных) характеристических функций игр совпадает с евклидовым пространством размерности 2^U-1. Аксиома аддитивности гласит, что если мы знаем функцию значения на некотором множестве игр, составляющих аддитивный базис для *G*, то мы можем определить значение для любой игры.

Набор игр, позволяющий достичь этого – это набор, состоящий из игр v_R, определенных для каждого подмножества *R* из *U*

$$v_R(S) = 1 \quad \text{если } R \subset S,$$
$$= 0 \quad \text{в остальных случаях.}$$

Любой игрок, не входящий в *R*, является нулевым игроком в этой игре, которую иногда называют «чистым торгом» или игрой «единодушия» между игроками в *R*, потому что все они должны договориться между собой, как разделить имеющееся богатство.

Поскольку все игроки в *R* симметричные, аксиома симметрии требует, чтобы $\phi_i(v_r) = \phi_j(v_r)$ для всех *i* и *j* в *R*, а аксиома нулевого игрока определяет $\phi_k(v_r) = 0$ для всех *k*, не в *R*, тогда аксиома эффективности позволяет сделать вывод, что $\phi_i(v_r) = 1/r$ для всех *i* в *R*, где *r*-количество игроков в *R*. (Для любой конечной коалиции *S* обозначим через *s* число игроков в *S*.) Таким образом, значение однозначно определено для всех игр вида v_r или для игр вида cv_r для любого числа *c* (где $cv_r(S)=c$, если $R \subset S$ и 0 в противном случае). (cv_r является супераддитивным, когда *c* неотрицателен.)

Однако, игры v_r формируют основу для множества всех игр, потому что есть 2^U-1 из них, по одному для каждого непустого подмножества *R* из *U*, и потому что они линейно независимы. Следовательно, любая игра v может быть записана как сумма игр в форме cv_r. Аксиома аддитивности подразумевает, что существует единственное значение, подчиняющееся аксиомам Шепли, определенным на пространстве всех игр.

Шепли показал, что это уникальное значение ϕ

$$\phi_i(v) = \sum_{S \subset N} \frac{(S-1)!(n-S)!}{n!} [v(S) - v(S-i)] \quad (1),$$

где *N* – любая конечная выборка, с |*N*| = *n*. Эта формула выражает значение Шепли для игрока *i* в игре *v* как взвешенную сумму членов вида [*v*(*S*) - *v* (S-*i*)], которые являются предельным вкладом *i*-го игрока в коалиции *S*.

Фактически, $\phi_i(v)$ можно интерпретировать как ожидаемый предельный вклад игрока *i*, где распределение коалиций возникает определенным образом [The Shapley value, 1988].

Использование вектора Шепли в регрессионном анализе осуществляется посредством задания характеристической функции через значение коэффициента детерминации [Михеенко и др., 2020].

Значение Шепли в интерпретация машинного обучения и критика метода

За счет значений Шепли возможно достаточно точно рассчитать вклад каждого игрока в возможные коалиции. Тем не менее некоторые авторы подчеркивали их вычислительную сложность, так как из уравнения 1 видно, что способ вычисления значения определяется только после вычисления всех возможных коалиций, которые для n числа игроков равны 2^U-1. Ряд исследователей разработали усовершенствованные алгоритмы для решения этой

сложной задачи. В первую очередь научные изыскания сконцентрированы в области создания эффективных конструкций с алгоритмами проектирования, позволяющими снизить вычислительную нагрузку [Landinez-Lamadrid и др., 2017]. Данная тенденция наряду с распространением формы объяснения моделей машинного обучения через оценку важности факторов, при которой каждому фактору приписывается некое значение, пропорциональное вкладу фактора в предсказание, становится определяющим в востребованности методов, основанных на применении значений Шепли [Luke Merrick и др., 2020].

В последнее время появилось несколько основанных на использовании значений Шепли методов сопоставления предсказания модели и ее входных факторов. Среди них выделяются SHAP, KernelSHAP, TreeSHAP, QII, andIME. В применении значений Шепли для моделей машинного обучения ключевую роль играет создание такой игры, в которой игроками будут факторы модели, а значение выигрыша будет равно предсказанию модели. Благодаря сильным аксиоматическим гарантиям, такой метод рассматривается, как, де-факто, подход к определению важности факторов, причем некоторые исследователи даже предполагают, что это может быть единственный метод, не противоречащий требованию «права на объяснение» в регламентах, таких как «Общий Регламент По Защите Данных» [Aas K. и др., 2019].

Вместе с этим, в исследовании Люка Меррика и Анкура Тали (Fiddler Labs, Palo Alto, USA, 2020) было определено, что при использовании значения Шепли при интерпретации моделей машинного обучения, в случае когда функция коррелирует с предсказанием модели на входных данных, формулировка игры приводит к ненулевой атрибуции функции независимо от того, влияет ли функция непосредственно на прогноз. В своем исследовании Люк Меррик и Анкура Тали установили присутствие случаев с противоестественным характером оценок важности факторов: «…метод SHAP определяет высокую важность фактора, который не имеет отношения к функционированию модели…».

Таким образом на практике интерпретация Шепли может давать «ложноположительные» результаты по важности предикторов.

Многоцикличная интерпретация модели

Алгоритмом, разрешающим проблему наличия вероятности ошибочной интерпретации важности фактора посредством использования векторов Шепли, может выступать разработанный нами выбор стабильной модели на основе метода Шепли (Selecting Stable Model SHAP (SSMS)).

Алгоритм SSMS основан на многоцикличном построении моделей, в которых посредством диапазональной доступности информации в обучающем наборе данных и комбинаций гиперпараметров обеспечивается различная точность, фиксирующаяся по завершению обучения. После построения поочередно вычисляются значения Шепли для объяснения предсказаний каждой модели.

Алгоритм формирует массив данных положений каждого предиктора на векторе важности факторов каждой модели с соответствующим значением ошибки данной модели. Сумма произведений данных показателей для соответствующего предиктора определяет его коэффициент устойчивости в рамках текущего моделирования в массиве всех заложенных атрибутов.

В общем виде цикл алгоритма имеет вид:

Входные параметры

Tx: набор входных предикторов для обучения.

Ty: набор значений целевой переменной для обучения.

Ex: набор значений входных предикторов для оценки точности модели.

Ey: набор значений целевой переменной для оценки точности модели.

Sp: стартовый размер доли Tx для обучения [0.0, 1.0].

AML – соответствующий ансамблевый алгоритм машинного обучения.

n – количество атрибутов выборки.

importances <- int A []=Tx.columns # *изначальная важность входных предикторов равна 0* при **Sp < 1.0** выполнять:

tx_cut <- Sp*Tx

ty_cut <- Sp*Ty

model <- AML(tx_cut, ty_cut) # *Тренировать AML-регрессор/классификатор на усеченном наборе данных*

wape <- WAPE(model, ex, ey) # *оценить ошибку модели по методике WAPE*

tree_explainer <- shap.TreeExplainer(model) # *подготовить объяснение для обученной модели*

shap_vals <- tree_explainer.shap_values(tx) # *получить вектора важности для каждого вектора входных данных*

weighted_factor_importances <- ∑ (|shap_vals|*|1.0-wape|)) # *получить результаты сумм произведений ординатных значений многомерного массива shap_vals и условного модуля точности соответствующей модели.*

//

Sp <- Sp + 0.01 # *шаг повышения диапазональной доступности информации в обучающем наборе данных*

// повторение цикла

sorted_importances <- СОРТИРОВКА(weighted_factor_importances) # *направление = «по убыванию»; в переменной sorted_importances будет сохранен отсортированный список факторов по их важности.*

ВОЗВРАТ sorted_importances # *вернуть отсортированный список вызывающей функции.*

 # В процессе выполнения цикла алгоритма больший вес получают атрибуции более точных моделей. Произведение массива атрибуций и точности мы определяем как «взвешенная важность факторов». Количество циклов может быть задано как сумма прогрессии количества атрибутов (n), так и пользователем в зависимости от задачи.

SSMS и устойчивость моделей

В целях проверки работы алгоритма и его воздействия на устойчивость моделей был сгенерирован тестовый набор данных, сформированный на 50 % из случайных не связанных величин, другая половина DataSet функционально связана с целевой переменной угловыми коэффициентами, заданными с экспотенциональным снижением. Связанные предикторы заданы случайными величинами в диапазоне аналогичном генерации несвязанных переменных (в целях усиления воздействия «шума»).

Алгоритм SSMS был построен в языковой среде Python и применен к наиболее распространённым и популярным алгоритмам машинного обучения – классу ансамблевых алгоритмов [Багутдинов и др. 2020; Guolin Ke и др. 2017]:

DecisionTreeRegressor (DTR) – алгоритм дерева решений.

BaggingRegressor (BR) – алгоритм бэггинга – ансамблевый мета-алгоритм, объединяющий предсказания из нескольких деревьев решений с помощью механизма «голосования большинством голосов» [Simske, 2020].

RandomForestRegressor (RFR) – алгоритм случайного леса – алгоритм на основе бэггинга, в котором подмножество объектов выбирается случайным образом для построения леса или ансамбля деревьев решений [Ghasemi и др. 2013].

XGBoost (XGB) – алгоритм машинного обучения, основанный на дереве поиска решений и использующий фреймворк градиентного бустинга. Разработан в рамках исследовательской деятельности Вашингтонского Университета, опубликован на конференции SIGKDD в 2016 году. Системно оптимизирован в параллелизации циклов выполнения алгоритмов бустинга и использовании параметра глубины вместо критерия

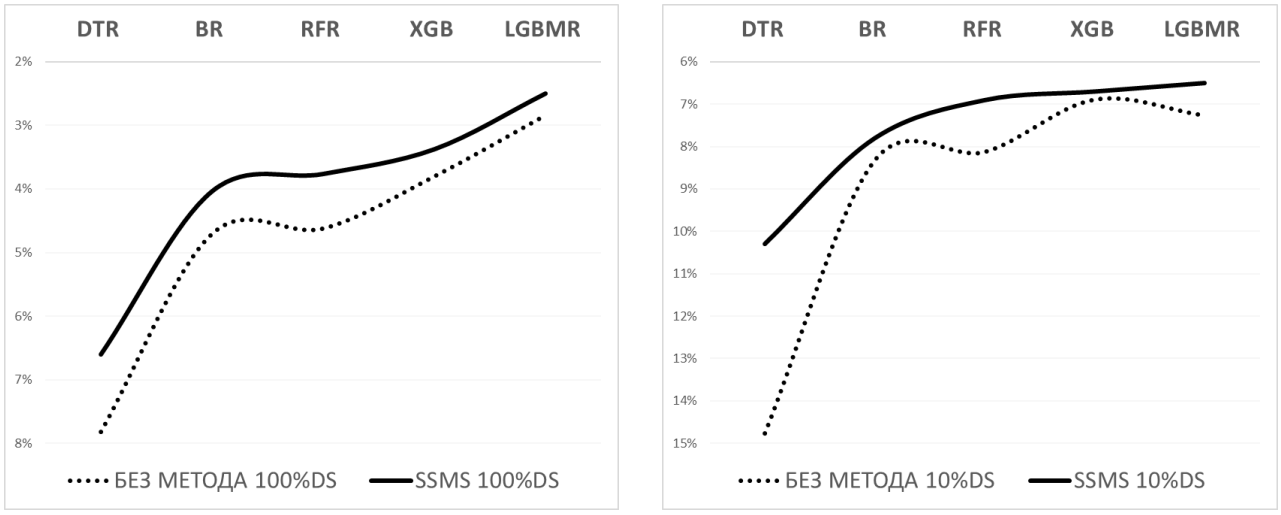
отрицательной потери для отсечения ветвей деревьев [Tianqi Chen и др., 2016]. XGBoost использует регуляризацию, схожую с ElasticNet (Elastic net regularization) – с целью исключения переобучения штрафуются сложные модели, использующие регуляризацию LASSO (L1) [Bochkarev и др. 2018] и Ridge-регуляризацию (L2) [Hoerl, 1987]. При работе со взвешенным DataSet оптимальные точки разделения в алгоритме находятся посредством метода взвешенных квантилей.

LightGBM (LGBMR) – алгоритм, разработанный в 2016 году исследовательской группой компании Майкрософт (Microsoft Research), опубликован в журнале «Достижения в области нейронных систем обработки информации» (Advances in Neural Information Processing Systems) в 2017 году. Алгоритм имеет доказанную повышенную скорость реализации градиентного бустинга (относительно других алгоритмов, в т. ч. XGB) благодаря своему уникальному методу построения деревьев и оптимизации памяти и вычислений на основе гистограмм.

Устойчивость к снижению данных

Для проверки воздействия алгоритма SSMS на точность прогноза при снижении набора данных были проведены замеры взвешенной абсолютной процентной ошибки прогнозирования (WAPE) при обучении на полном DataSet и на усеченном наборе, составляющем 10 % от выборки.

При применении SSMS определяется рост качества прогноза по всем анализируемым алгоритмам (рис. 1).



А Б

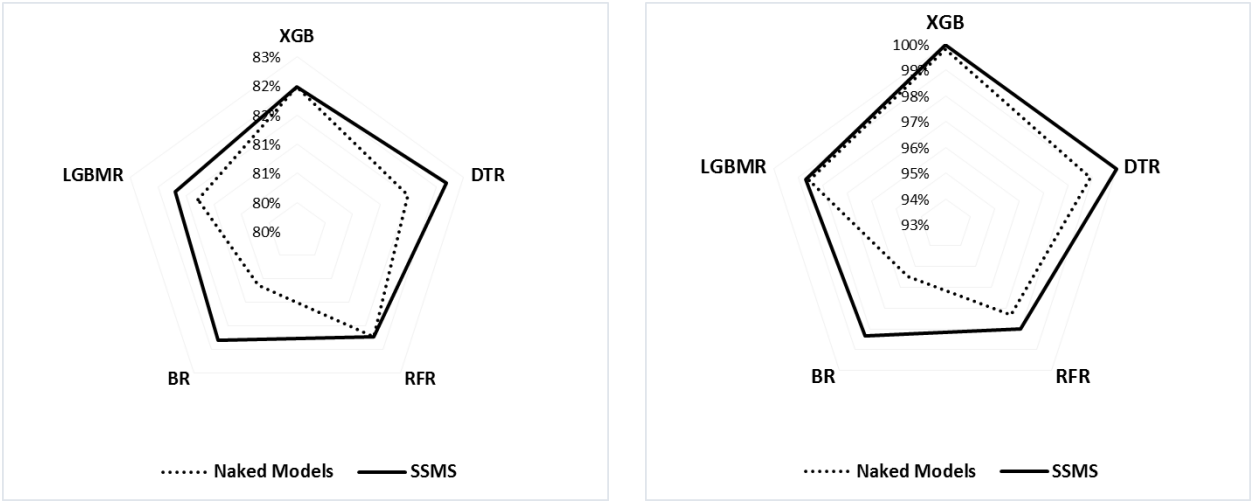
Рис. 1. WAPE при применении SSMS и без него по основным ансамблевым алгоритмам при обучении на полном тестовом наборе данных (А) и усечённом наборе (Б)
Fig. 1. WAPE of main ensemble algorithms trained on full dataset (A) and on sliced dataset (B) with SSMS and without it

SSMS улучшает показатели точности при снижении набора данных даже на улучшенных и модернизированных ансамблевых алгоритмах. Так высокоэффективный «древесный» алгоритм LightGBM имеет внедренные методы GOSS (Gradient-based One-Side Sampling – определяющий исключение значительной части данных с небольшими градиентами) и EFB (Exclusive Feature Bundling – связывающий взаимоисключающие функции) [Guolin Ke и др. 2017], однако и в нем реализован потенциал снижения WAPE, посредством применения SSMS.

Устойчивость к искажению предикторов

Имитация дестабилизации вводных параметров достигалась путем попеременного искажены предикторов на случайную величину в диапазоне всех переменных тестового набора.

При применении SSMS определяется рост качества прогноза по большинству анализируемым алгоритмам (рис. 2)



А Б

Рис. 2. Точность прогноза моделей при искажении предикторов, функционально связанных с целевой переменной (А) и искажении случайных величин, добавленных в набор для обучения (Б)
Fig. 2. Forecast precision of models upon distorting functionally relevant features (A) and distorting random features added into the training set (B)

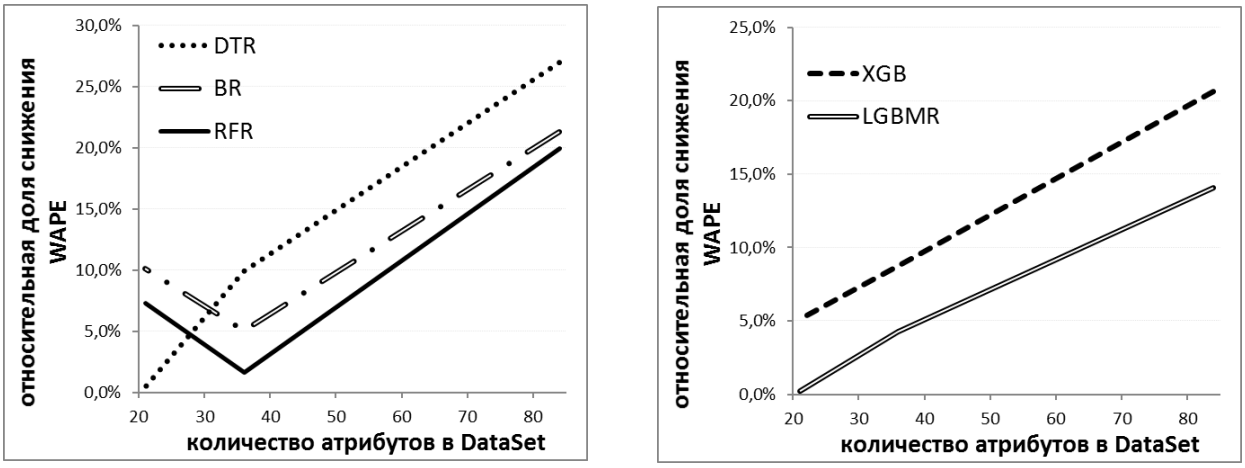
В случае искажения функционально связанных переменных (истинных предикторов) SSMS повышает точность прогноза алгоритмов дерева решений, бэггинга и LightGBM, неизменной остается точность моделей, построенных по алгоритмам случайного леса и градиентного бустинга (XGB). При искажении несвязанных переменных (шума) применение SSMS не оказывает выраженного положительного воздействия на алгоритмы XGB и LightGBM, по остальным моделям позволяет повысить точность прогноза.

Проверка алгоритма в несинтезированных наборах

В целях анализа работы метода SSMS с наборами данных различной размерности мы провели эксперимент с популярными DataSet для машинного обучения в задачи регрессии: FIFA 19 complete player dataset (размерность: 84 атрибута х 18000 записей); Cancer Linear Regression Model Tutorial (размерность: 36 х 3048); Life Expectancy (WHO) (размерность: 21 х 2938) [Конкурсная платформа, 2020; Ресурс открытых данных, 2020].

В проведенных экспериментах наблюдается выраженное сокращение WAPE при росте атрибутной размерности выборки и соответствующие снижение результативности SSMS при ограниченном количестве предикторов (см. рис.3).

Снижение WAPE достигает до четверти от первоначальных величин на наборе данных со значительным количеством атрибутов. SSMS позволяет отбирать модели, включающие исключительно важные предикторы, игнорируя шум, что и определяет наблюдаемую зависимость. Аналогичные результаты фиксировались и при решении задач классификации.



А Б

Рис. 3 – относительное снижение WAPE с применением SSMS (к значениям WAPE моделей построенных без SSMS) по «классическим» (А) и модернизированным (Б) ансамблевым алгоритмам
Fig. 3. Relative (to WAPE of models trained without SSMS) decrease of WAPE with SSMS on "classic" (A)and innovative ensemble algorithms (B)

Заключение

Предложенный алгоритм выбора стабильной модели на основе значений Шепли (SSMS) позволяет повышать показатели точности при снижении набора данных на классических и модернизированных ансамблевых алгоритмах.

SSMS стабилизирует показатели точности моделей при искажении предикторов и шумов в большинстве ансамблевых алгоритмов, в том числе в высокоэффективном и высокоскоростном алгоритме LightGBM.

С ростом размерности выборки по количеству атрибутов растет эффективность применения предложенного алгоритма, определяемая процедурой выбора стабильно важных по значению Шепли предикторов.

Ограничением использования SSMS является повышенная ресурсоемкость метода относительно моно-использования ансамблевых алгоритмов при динамическом обновлении и переобучении. В случае приоритезации точности моделей над ресурсоемкостью, ограниченном по количеству записей DataSet, либо в случае высокой варьируемости ключевых предикторов, SSMS может выступать в качестве инструмента повышения эффективности ансамблевых алгоритмов машинного обучения.

Список литературы

1. Багутдинов Р.А., Саргсян Н.А., Красноплахтыч М.А. 2020. Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 47 (4): 792–802.

2. Конкурсная платформа по исследованию данных Kaggle Machine Learning Competition Platform (Google). 2020. [Электронный ресурс]. URL: https://www.kaggle.com/datasets (Дата обращения 04.10.2020).

3. Михеенко А.М., Савич Д.С. 2020. Вестник Балтийского федерального университета им. И. Канта. Сер.: Физико-математические и технические науки. № 2. 84–94.

4. Ресурс данных для машинного обучения Data.world. 2020. [Электронный ресурс]. URL:https://data.world/ (Дата обращения 26.11.2020).

5. Aas K., Jullum M., LØland A.2021. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence. 298:103502. DOI10.1016/ j.artint.2021.103502.

6. Alvin E. Roth. 1988. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press. ISBN0-521-36177-X.

7. Bochkarev V., Tyurin V., Savinkov A., Gizatullin B. 2018. Application of the LASSO algorithm for fitting the multiexponential data of the NMR relaxometry. Journal of Physics Conference Series. 1141(1):012148. DOI10.1088/1742-6596/1141/1/012148.

8. Chen T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754. DOI 10.1145/2939672.2939785.

9. Ghasemi J.B. Tavakoli H. 2013. Application of Random Forest Regression to Spectral Multivariate Calibration. Analytical Methods. 5 (7):1863–1871. DOI10.1039/C3AY26338J.

10. Hoerl R. 1987.The Application of Ridge Techniques to Mixture Data: Ridge Analysis. Technometrics. 29 (2):161–172. DOI10.1080/00401706.1987.10488207.

11. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS).

12. Landinez-Lamadrid D.C., Ramirez-Ríos D.G., Neira Rodado D., Parra Negrete K. and Combita Niño J.P. 2017. Shapley Value: its algorithms and application to supply chains. INGE CUC, 13 (1): 61–69.

13. Mason Ch. H., Perreault Jr. W.D. 1991.Collinearity, power, and interpretation of multiple regression analysis. Journal of Marketing Research. 28: 268–280.

14. Merrick L. and Taly A. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. Fiddler Labs, Palo Alto, USA. arXiv:1909.08128. DOI10.1007/978-3-030-57321-8_2.

15. Simske S. J. 2015.The rationale for ensemble and meta-algorithmic architectures in signal and information processing. APSIPA Transactions on Signal and Information Processing. 4: 1–9. DOI10.1017/ATSIP.2015.10.

References

1. Baguydinov R.A., Sargsyan N.A., Krasnoplahtysh M.A. 2020. Analitika, instrumenty i intellektualny analiz bolshih raznorodnyh i raznomasshtabnyh dannyh. [Analytics, tools, and intelligent analysis of large heterogeneous and multi-scale data]. Scientific Bulletin of the Belgorod State University. Series: Economics. Computer science. 47 (4): 792–802.

2. Konkursnaya platforma po issledovaniu dannyh Kaggle Machine Learning Competition Platform (Google). [Competitive Data Research Platform Kaggle Machine Learning Competition Platform (Google)] 2020. [Electronic resource]. URL: https://www.kaggle.com/datasets (Date of access 04.10.2020).

3. Mikheenko A.M., Savich D.S. 2020. Vestnik Baltiyskogo Federalnogo Universiteta im. Kanta [Bulletin of the Baltic Federal University named after I. Kant.] Ser.: Physical-mathematical and technical sciences. № 2. 84–94.

4. Machine learning training data source Data.world. 2020. [Electronic resource]. URL:https://data.world/ (Date of access 26.11.2020).

5. Aas K., Jullum M., LØland A.2021. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence. 298:103502. DOI10.1016/j.artint.2021.103502.

6. Alvin E. Roth. 1988. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press. ISBN0-521-36177-X.

7. Bochkarev V., Tyurin V., Savinkov A., Gizatullin B. 2018. Application of the LASSO algorithm for fitting the multiexponential data of the NMR relaxometry. Journal of Physics Conference Series. 1141(1):012148. DOI10.1088/1742-6596/1141/1/012148.

8. Chen T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754. DOI 10.1145/2939672.2939785.

9. Ghasemi J. B. Tavakoli H. 2013. Application of Random Forest Regression to Spectral Multivariate Calibration. Analytical Methods. 5(7):1863–1871. DOI10.1039/C3AY26338J.

10. Hoerl R. 1987.The Application of Ridge Techniques to Mixture Data: Ridge Analysis. Technometrics. 29 (2):161–172. DOI10.1080/00401706.1987.10488207.

11. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS).

12. Landinez-Lamadrid D.C., Ramirez-Ríos D. G., Neira Rodado D., Parra Negrete K. and Combita Niño J.P. 2017. Shapley Value: its algorithms and application to supply chains. INGE CUC, 13 (1): 61–69.

13. Mason Ch. H., Perreault Jr. W.D. 1991.Collinearity, power, and interpretation of multiple regression analysis. Journal of Marketing Research. Vol. 28. 268–280.

14. Merrick L. and Taly A. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. Fiddler Labs, Palo Alto, USA. arXiv:1909.08128. DOI10.1007/978-3-030-57321-8_2.

15. Simske S. J. 2015.The rationale for ensemble and meta-algorithmic architectures in signal and information processing. APSIPA Transactions on Signal and Information Processing. 4: 1–9. DOI10.1017/ATSIP.2015.10.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Воробьев Александр Викторович, аспирант кафедры ПОАИС Курского государственного университета, г. Курск, Россия

INFORMATION ABOUT THE AUTHOR

Alexander V. Vorobyev, Postgraduate Cathedra of SISA, Kursk State University, Kursk, Russia