УДК 519.248:658.518.5 doi: 10.15827/0236-235X.142.189-195 2023. Т. 36. № 2. С. 189–195

Сравнительный анализ методов построения математических моделей функционирования объекта с применением машинного обучения

В.Н. Ковальногов
В.В. Шеркунов
Хуссейн Мохамед Хуссейн
В.Н. Клячкин

Ссылка для цитирования
Ковальногов В.Н., Шеркунов В.В., Хуссейн Мохамед Хуссейн, Клячкин В.Н. Сравнительный анализ методов построения математических моделей функционирования объекта с применением машинного обучения // Программные продукты и системы. 2023. Т. 36. № 2. С. 189–195. doi: 10.15827/0236-235X.142.189-195
Информация о статье
Поступила в редакцию: 27.10.2022 После доработки: 30.01.2023 Принята к публикации: 14.02.2023

Аннотация. Предметом данного исследования является технический объект, работа которого определяется множеством факторов, а качество функционирования характеризуется некоторым показателем. Требуется построить математическую модель, связывающую этот показатель со значениями факторов. В качестве примера исследуется влияние различных факторов на эффективность работы горелочных устройств (нагрузки, расхода воздуха, метана и биогаза, составов топлива и окислителя и других). Эффективность (качество функционирования) горелочного устройства оценивается по температуре дымовых газов. Задача решается методами машинного обучения, поскольку классические методы регрессионного анализа показали недостаточную точность. В настоящей статье исследуется эффективность метода опорных векторов, случайного леса и бустинга деревьев решений. Для численных расчетов использована локализованная версия 13.3 системы Statistica. Все три подхода машинного обучения показали существенное повышение точности модели на тестовой выборке. Наилучшие результаты в рассматриваемом примере дал метод бустинга деревьев решений. Рекомендуемая технология построения модели, обеспечивающая необходимую точность прогнозирования, сводится вначале к апробации классического регрессионного анализа (если полученная модель обеспечит необходимую точность, то она предпочтительна с точки зрения ее интерпретируемости). При недостаточной точности используются три рассмотренных метода машинного обучения, вместе с тем важен подбор параметров каждого из них, который, с одной стороны, обеспечивал бы необходимую точность, а с другой – не приводил бы к переобучению модели. Полученная модель может быть использована для оценки влияния различных факторов на эффективность работы технического объекта, а также для прогнозирования качества его функционирования, в частности, температуры дымовых газов.

Ключевые слова: регрессионная модель, мультиколлинеарность, метод опорных векторов, случайный лес, бустинг деревьев решений
Благодарности. Исследования поддержаны грантом Президента Российской Федерации, проект НШ-28.2022.4

Рассматривается технический объект, работа которого определяется множеством p факторов X_j, а качество функционирования характеризуется показателем Y. Известны результаты наблюдений за работой объекта. Требуется построить математическую модель, связывающую показатель Y со значениями факторов X_j.

Это стандартная задача построения множественной регрессии, решение которой при определенных условиях можно использовать для прогнозирования значений – откликов Y по заданному набору показателей X_j. Проблема состоит в том, что далеко не всегда такую модель можно корректно построить: она может оказаться незначимой или при значимости по критерию Фишера недостаточно качественной для прогнозирования вследствие низкого коэффициента детерминации – квадрата коэффициента корреляции между опытными и прогнозируемыми значениями (показывает, какая доля дисперсии отклика может быть объяснена рассматриваемыми факторами) [1].

В этом случае более эффективным может быть применение нейронных сетей. Известно, что глубокое обучение сетей приводит к существенному повышению качества построенной модели. Однако для глубокого обучения необходим достаточно большой объем выборочных данных, что для реальных технических объектов, как правило, получить невозможно: обычно выборки имеют объем в несколько десятков или сотен наблюдений [2, 3].

В настоящей статье в качестве примера исследуется влияние различных факторов на эффективность работы горелочных устройств (нагрузки, расхода воздуха, метана и биогаза, составов топлива и окислителя и других). Эффективность горелочного устройства Y оценивается по температуре дымовых газов.

Для численных расчетов использовалась локализованная версия 13.3 системы Statistica.

Как правило, решение задач машинного обучения осуществляется путем разработки соответствующей программы на языке программирования Python, в котором есть множество уже отлаженных конструкторов для задач классификации и регрессии, а также метрик для оценки качества полученных моделей. В частности, аналогичная задача в статье [4] решалась с помощью такой программы другим методом – путем разделения состояний горелочного устройства на оптимальное, удовлетворительное и неудовлетворительное (мультиклассовая классификация).

Вместе с тем при наличии в организации системы Statistica нужный результат может быть получен гораздо оперативнее. Эта система разработана американской компанией, адаптирована к отечественной практике и является самой распространенной статистической системой в России.

Постановка задачи

Эффективность функционирования рассматриваемого горелочного устройства, по мнению экспертов, определялась 20 факторами. Три пары показателей оказались связанными линейными зависимостями, таким образом, три фактора были исключены из рассмотрения (табл. 1).

Также исследовалось наличие корреляционных связей между оставшимися 17 показателями. Сильная корреляция (выборочный коэффициент корреляции $r > 0{,}9$) имеет место между парами показателей X_4–X_5, X_4–X_9, X_5–X_9, X_6–X_7, X_6–X_{11}. Однако, по предложению экспертов, все эти показатели были учтены в расчетах.

Наличие выбросов в исходных данных оценивалось приближенно по диаграммам рассеяния между парами показателей. Всего из 309 наблюдений обнаружено 9 выбросов. Таким образом, число наблюдений равно 300.

По этим данным строилась регрессионная модель с учетом ее мультиколлинеарности (наличия сильных корреляций между факторами). Использовалась гребневая регрессия. При этом незначимые по критерию Стьюдента факторы отсеивались: использовался алгоритм пошаговой регрессии.

Этот алгоритм одновременно с гребневой регрессией реализован в системе Statistica. Результаты расчета показаны в таблице 2. Для обучения модели использованы 240 наблюдений из 300: 60 наблюдений оставлены для последующего тестирования, чтобы исключить переобучение модели. Из 17 факторов значимыми оказались только четыре: X_1 (нагрузка), X_{14} (температура топлива), X_{16} (размер сетки) и X_{17} (коэффициент избытка воздуха):

$$Y = 401{,}67 + 0{,}0376X_1 + 1{,}2883X_{14} + 27{,}1875X_{16} - 45{,}773X_{17}.$$

Параметр гребневой регрессии $\lambda = 0{,}001$ подобран из условия обеспечения максимума коэффициента детерминации. Модель оказалась значимой по F-критерию Фишера (вероятность ошибки $p < 0{,}05$), все входящие в модель факторы значимы по t-критерию Стьюдента (вероятности ошибок $p < 0{,}05$), при этом коэффициент детерминации R^2 оказался равным 0,37, что является недопустимо низким значением.

Таблица 1

Показатели работы горелочного устройства

Table 1

Burner performance indicators

Обозначение	Наименование, единица измерения	Значение	
		Минимальное	Максимальное
X_1	Нагрузка, т/ч	170	500
X_2	Расход воздуха, $м^3/ч$	5044	59 719
X_3	Расход метана, $м^3/ч$	0	5 375
X_4	Расход биогаза, $м^3/ч$	0	5 000
	Состав топлива, %:		
X_5	CH_4	30	98
X_6	C_2H_6	0	15
X_7	C_3H_8	0	9
X_8	CO_2	0	32
X_9	N_2	0	1,05
	Состав окислителя, %:		
X_{10}	O_2	0,179	0,232
X_{11}	N_2	0,750	0,768
X_{12}	CO_2	0	0,023
X_{13}	Температура воздуха, К	446	533
X_{14}	Температура топлива, К	10	25
X_{15}	Угол наклона лопаток завихрителя, град.	0	50
X_{16}	Размер сетки, м	0,1	1
X_{17}	Коэффициент избытка воздуха	0,88	1,5
Y	Температура дымовых газов, К	348	412

Таблица 2

Результаты расчета регрессии

Table 2

Regression calculation results

Фактор	Коэффициент B	Стандартная ошибка B	*t*(235)	Значение *p*
Св. член	401,67	14,16	28,36	0,0000
X_1	0,0376	0,00994	3,77	0,0002
X_{14}	1,2883	0,31684	4,07	0,0001
X_{16}	27,1875	4,16803	6,52	0,0000
X_{17}	-45,773	11,46809	3,99	0,0001

Примечание. Гребневая регрессия для зависимой переменной *Y*, наблюдений – 240, $\lambda = 0{,}001$, $R^2 = 0{,}37$; $F(4,235) = 33{,}85$; $p < 0{,}000$; стандартная ошибка оценки 12,94.

Именно это обстоятельство и привело к поиску других методов построения модели. Обзор методов построения регрессий с использованием машинного обучения выявил возможность использования трех подходов для получения наиболее качественных моделей: метода опорных векторов [5–7], случайного леса [8, 9] и бустинга деревьев решений [10, 11].

Эти методы использовались для решения различных задач построения регрессий, например, для прогнозирования работы системы водоочистки, при вибромониторинге гидроагрегата, в задаче оценки стабильности функционирования газотурбинного двигателя и других. При этом выявлено, что ни один из методов не гарантирует достаточно качественное построение модели (за исключением глубокого обучения нейросетей, но, как известно, этот метод связан с требованием слишком большого объема наблюдений). В зависимости от конкретного набора исходных данных возможны как недостаточно высокая точность модели, так и ее переобучение.

Цель исследования – разработать технологию построения регрессионной модели, обеспечивающую необходимую точность прогнозирования показателя эффективности функционирования технического объекта, путем выбора соответствующего метода обучения и оценки его параметров.

Метод опорных векторов

Данный метод основан на разделении объектов гиперплоскостью способом, максимизирующим ширину разделяющей полосы – зазор между опорными векторами. Для линейно неразделимых данных используют различные варианты функции ядра. Программа позволяет выбрать тип ядра линейный, полиномиальный, сигмоидный и радиальную базисную функцию. В рассматриваемой задаче опробованы различные типы ядер и выбрана радиальная базисная функция с параметром гамма, равным 0,0588 и обеспечивающим наилучшие предсказанные значения. При необходимости параметры могут быть уточнены с применением кросс-валидации.

На рисунке 1 показаны опытные и предсказанные значения отклика для тестовой части выборки.

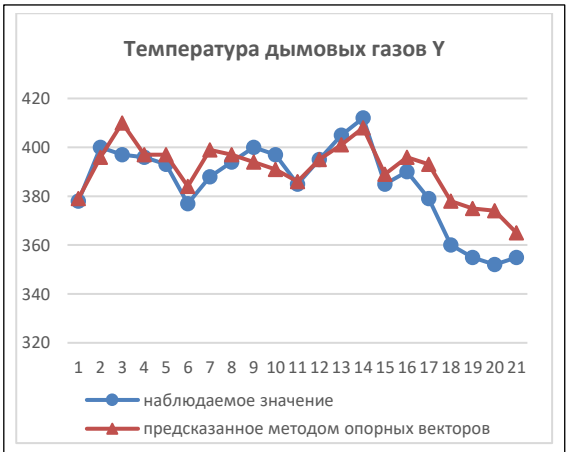

Рис. 1. Предсказанные значения по тестовой выборке с помощью метода опорных векторов

Fig. 1. Predicted values from the test sample using the support vector machine

По этим данным подсчитывались две характеристики качества построенной модели:

– средняя абсолютная процентная ошибка (MAPE):

$$MAPE = \frac{1}{n_T}\sum_{i=1}^{n_T}\frac{|\tilde{y}_i - y_i|}{|y_i|}100\,\%,$$

где n_T – объем тестовой выборки; y_i – опытное значение отклика; $\tilde{y}_i$ – прогнозируемое значение по построенной модели;

– корень из средней квадратичной ошибки (RMSE):

$$RMSE = \sqrt{\frac{1}{n_T}\sum_{i=1}^{n_T}(\tilde{y}_i - y_i)^2}.$$

Для данных по рисунку 1 определим $MAPE = 2{,}09$ %, $RMSE = 10{,}2$.

Полученные значения будем далее сравнивать с соответствующими характеристиками моделей, построенных другими методами.

Случайный лес

Алгоритм сочетает в себе случайный выбор с возвращением и метод случайных подпро-

странств. Он состоит из множества независимых деревьев решений, при этом используются случайная выборка наблюдений из обучающего набора и случайный набор показателей при принятии решений о разбиении узлов. Случайный лес применяется для решения задач классификации, регрессии и кластеризации.

Метод имеет высокую точность предсказания, нечувствителен к монотонным преобразованиям значений показателей, редко переобучается: добавление деревьев почти всегда только улучшает композицию, но после достижения определенного количества деревьев кривая обучения выходит на асимптоту. К недостаткам относят то, что в отличие от одного дерева результаты случайного леса сложнее интерпретировать; кроме того, требуется много памяти для хранения модели вследствие большого размера получающихся моделей.

Программа Statistica показывает ход процесса обучения с помощью случайного леса, построенное дерево (рис. 2), столбчатую диаграмму важности факторов по степени их влияния на отклик (рис. 3), а также прогнозируемые значения отклика на тестовой выборке.

С учетом прогнозируемых этим методом значений получим: средняя абсолютная процентная ошибка *MAPE* = 2,25 %, корень из средней квадратичной ошибки *RMSE* = 10,8.

Очевидно, что в рассматриваемой задаче точность прогнозирования методом случайного леса ниже, чем методом опорных векторов.

Бустинг деревьев решений

В ходе обучения случайного леса каждый базовый алгоритм строится независимо от остальных. В бустинге используется идея последовательного построения линейной комбинации алгоритмов. Каждый следующий алгоритм старается уменьшить ошибку текущего ансамбля.

Бустинг, использующий деревья решений в качестве базовых алгоритмов, называется градиентным бустингом над решающими

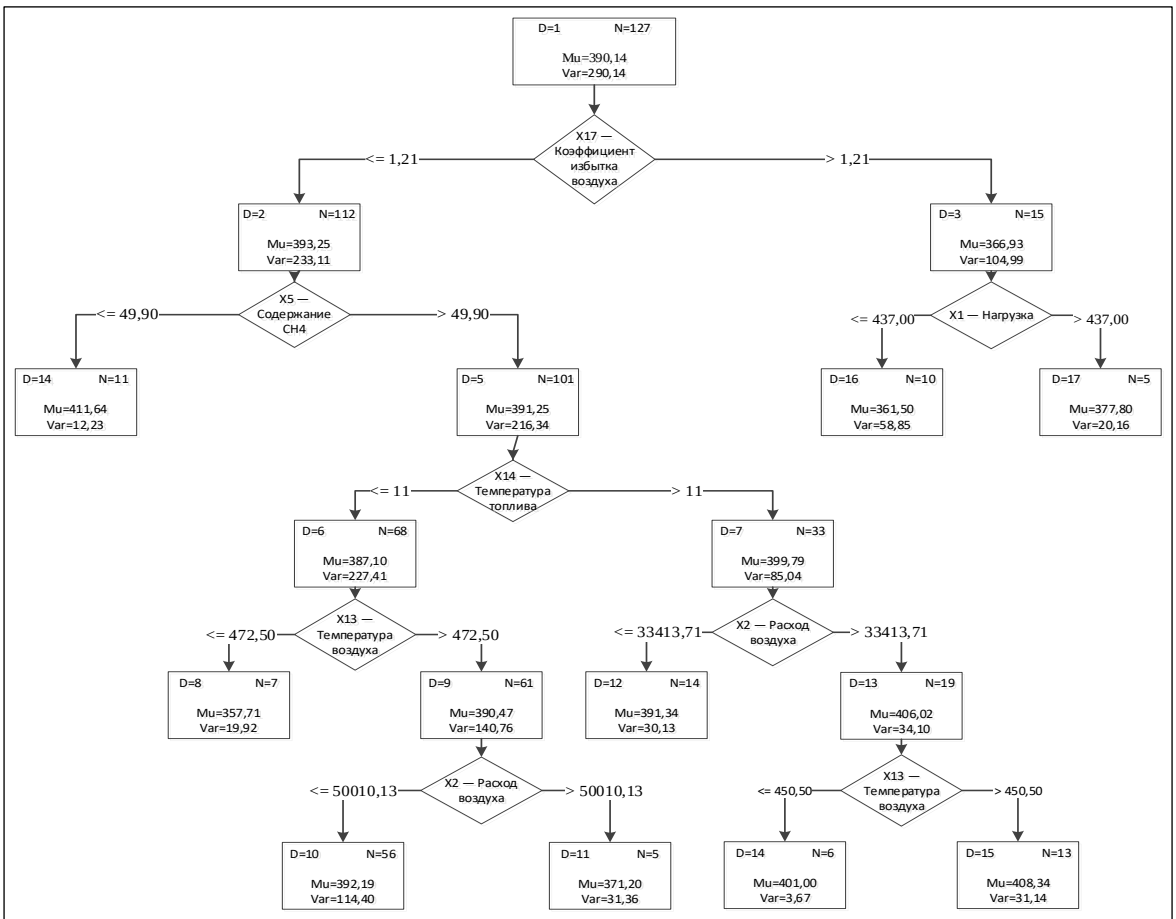

Рис. 2. Дерево, построенное методом случайного леса

Fig. 2. A tree built using the random forest method

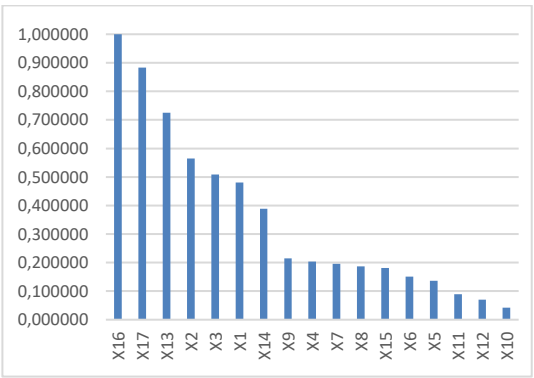

Рис. 3. Столбиковая диаграмма значимости факторов

Fig. 3. A bar chart of the factor significance

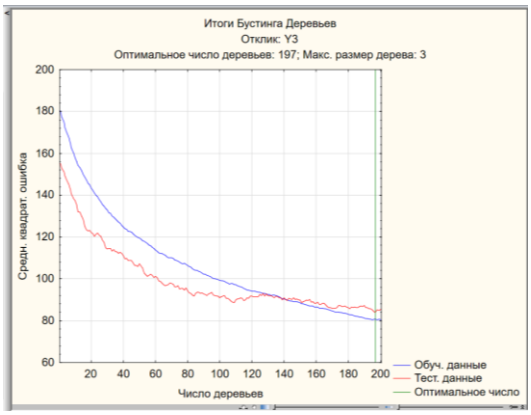

Рис. 4. Ход процесса обучения при градиентном бустинге деревьев решений

Fig. 4. The course of the learning process with gradient boosting of decision trees

деревьями. Если обучить одно дерево, то качество модели, скорее всего, будет низким. Однако о построенном дереве известно, на каких объектах оно давало точные предсказания, а на каких ошибалось. Таким образом, если вторая модель научится предсказывать разницу между реальным значением и ответом первой, то это позволит уменьшить ошибку композиции. Процесс продолжается, пока ошибка не минимизируется.

Настройки бустинга в системе Statistica были установлены по результатам предварительных испытаний: количество шагов – 200, минимальное число наблюдений – 7, максимальное количество уровней – 10. На рисунке 4 показан ход процесса обучения: синяя линия – средняя квадратичная ошибка на обучающей выборке, красная – на тестовой. Оптимальное число деревьев оказалось равным 197.

Программа, как и другие методы, выводит прогнозируемые значения отклика по тестовой выборке. С их учетом *MAPE* = 1,93 %, *RMSE* = 9,2.

Видно, что точность прогнозирования при использовании бустинга оказалась выше, чем двумя ранее рассмотренными методами, по обоим критериям.

Заключение

Построение математической модели функционирования технического устройства по результатам опытной эксплуатации методами регрессионного анализа по ограниченному объему наблюдений не всегда обеспечивает необходимое качество построенных моделей. Для повышения точности прогнозирования может оказаться полезным применение методов машинного обучения. Все три рассмотренных в статье подхода обучения с помощью метода опорных векторов, случайного леса и бустинга деревьев решений показали существенное повышение точности модели на тестовой выборке. Наилучшие результаты в рассматриваемом примере дал метод бустинга деревьев решений.

Таким образом, рекомендуемая технология построения математической модели, обеспечивающая необходимую точность прогнозирования показателя эффективности функционирования технического объекта, сводится к апробации вначале классического регрессионного анализа (если полученная модель обеспечит необходимую точность, то она предпочтительна с точки зрения ее интерпретируемости). При недостаточной точности используются три рассмотренных метода машинного обучения, при этом следует обратить внимание на необходимость подбора параметров каждого из методов, которые, с одной стороны, обеспечивали бы требуемую точность, с другой, не приводили бы к переобучению модели.

Список литературы

1. Клячкин В.Н., Крашенинников В.Р., Кувайскова Ю.Е. Прогнозирование и диагностика стабильности функционирования технических объектов. М.: РУСАЙНС, 2020. 200 с.
2. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение; [пер. с англ.]. М.: ДМК Пресс, 2018. 652 с.
3. Hanin B. Universal function approximation by deep neural nets with bounded width and ReLU activations. Mathematics, 2019, no. 7, art. 992. doi: 10.3390/math7100992.

4. Kovalnogov V., Fedorov R., Klyachkin V., Generalov D., Kuvayskova Y., Busygin S. Applying the random forest method to improve burner efficiency. Mathematics, 2022, no. 10, art. 2143. doi: 10.3390/math10122143.

5. Bavazeer S.A., Baakeem S.S., Mohamad A.A. A New radial basis approach based on Hermite expansion with respect to the shape parameter. Mathematics, 2019, no. 7, art. 979. doi: 10.3390/math7100979.

6. Sun X., Du P., Wang X., Ma P. Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. J. of the American Statistical Association, 2018, vol. 113, no. 524, pp. 1601–1611. doi: 10.1080/01621459.2017.1356320.

7. Pedregosa F., Bach F., Gramfort A. On the consistency of ordinal regression methods. J. of Machine Learning Research, 2017, no. 18, pp. 1–35.

8. Chen R., Paschalidis I. A robust learning approach for regression models based on distributionally robust optimization. J. of Machine Learning Research, 2018, no. 19, pp. 1–48.

9. Devijver E., Perthame E. Prediction regions through inverse regression. J. of Machine Learning Research, 2020, no. 21, pp. 1–24.

10. Генрихов И.Е., Дюкова Е.В., Журавлёв В.И. Построение и исследование полных решающих деревьев для задачи восстановления регрессии в случае вещественнозначной информации // Машинное обучение и анализ данных. 2017. Т. 3. № 2. С. 107–118. doi: 10.21469/22233792.3.2.02.

11. Park Ch. Jump gaussian process model for estimating piecewise continuous regression functions. J. of Machine Learning Research, 2022, no. 23, pp. 1–37.

Software & Systems doi: 10.15827/0236-235X.142.189-195 2023, vol. 36, no. 2, pp. 189–195

A comparative analysis of methods for constructing mathematical models of object functioning using machine learning

Vladislav N. Kovalnogov
Vyacheslav V. Sherkunov
Hussein Mohamed Hussein
Vladimir N. Klyachkin

For citation
Kovalnogov, V.N., Sherkunov, V.V., Hussein Mohamed Hussein, Klyachkin, V.N. (2023) 'A comparative analysis of methods for constructing mathematical models of object functioning using machine learning', *Software & Systems*, 36(2), pp. 189–195 (in Russ.). doi: 10.15827/0236-235X.142.189-195

Article info
Received: 27.10.2022 After revision: 30.01.2023 Accepted: 14.02.2023

Abstract. The subject of the study is a technical object; its work is determined by many factors, its performance is characterized by some indicator. It is necessary to build a mathematical model that connects this indicator with the values of factors. As an example, the article examines the influence of various factors on the efficiency of burner devices (load, air consumption, methane and biogas, fuel and oxidizer compositions, and others). The efficiency (performance) of the burner device is assessed by the temperature of the flue gases. The problem is solved by machine learning methods, since classical regression analysis methods showed insufficient accuracy. The article explores the effectiveness of the following approaches: the support vector method, random foresting and decision tree boosting. The authors used a localized version 13.3 of the Statistica system for numerical calculations. All three machine learning approaches discussed in the paper have shown a significant increase in the model accuracy on the test sample. The method of boosting decision trees has shown the best results in this example. The recommended model construction technology that provides the necessary forecasting accuracy is first reduced to testing the classical regression analysis (if the resulting model provides the necessary accuracy, then it is preferable from the point of view of its interpretability). If the accuracy is insufficient, the three considered methods of machine learning are used. It this case, it is important to select the parameters of each of the methods, which, on the one hand, would provide the necessary accuracy, on the other hand, would not lead to model retraining. The resulting model can be used to assess the influence of various factors on the efficiency of the technical facility, as well as to predict its functioning quality (in particular in the considered example, to predict the temperature of flue gases).

Keywords: regression model, multicollinearity, support vector method, random forest, decision tree busting

Acknowledgements. The research was supported by a grant from the President of the Russian Federation, project NSh-28.2022.4

Reference List

1. Klyachkin, V.N., Krasheninnikov, V.R., Kuvajskova, Yu.E. (2020) *Forecasting and Diagnostics of the Stability of the Technical Object Functioning*, Moscow (in Russ.).

2. Goodfellow, I., Bengio, Y., Courville, A. (2016) *Deep Learning*, Cambridge, Massachusetts, MIT Press (Russ. ed.: (2018) Moscow).

3. Hanin, B. (2019) 'Universal function approximation by deep neural nets with bounded width and ReLU activations', *Mathematics*, (7), art. 992. doi: 10.3390/math7100992.

4. Kovalnogov, V., Fedorov, R., Klyachkin, V., Generalov, D., Kuvayskova, Y., Busygin, S. (2022) 'Applying the random forest method to improve burner efficiency', *Mathematics*, (10), art. 2143. doi: 10.3390/math10122143.

5. Bavazeer, S.A., Baakeem, S.S., Mohamad, A.A. (2019) 'A New radial basis approach based on Hermite expansion with respect to the shape parameter', *Mathematics*, (7), art. 979. doi: 10.3390/math7100979.

6. Sun, X., Du, P., Wang, X., Ma, P. (2018) 'Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework', *J. of the American Statistical Association*, 113(524), pp. 1601–1611. doi: 10.1080/01621459.2017.1356320.

7. Pedregosa, F., Bach, F., Gramfort, A. (2017) 'On the consistency of ordinal regression methods', *J. of Machine Learning Research*, (18), pp. 1–35.

8. Chen, R., Paschalidis, I. (2018) 'A robust learning approach for regression models based on distributionally robust optimization', *J. of Machine Learning Research*, (19), pp. 1–48.

9. Devijver, E., Perthame, E. (2020) 'Prediction regions through inverse regression', *J. of Machine Learning Research*, (21), pp. 1–24.

10. Genrikhov, I.E., Djukova, E.V., Zhuravlyov, V.I. (2017) 'Construction and investigation of full regression trees in regression restoration problem in the case of real-valued information', *Machine Learning and Data Analysis*, 3(2), pp. 107–118 (in Russ.).

11. Park, Ch. (2022) 'Jump gaussian process model for estimating piecewise continuous regression functions', *J. of Machine Learning Research*, (23), pp. 1–37.

Авторы

Ковальногов Владислав Николаевич [1], д.т.н., зав. кафедрой тепловой и топливной энергетики, kvn@ulstu.ru

Шеркунов Вячеслав Владимирович [1], аспирант, v.sherkunov@ulstu.ru

Хуссейн Мохамед Хуссейн [1], аспирант, mohammedab634@gmail.com

Клячкин Владимир Николаевич [1], д.т.н., профессор кафедры прикладной математики и информатики, v_kl@mail.ru

[1] Ульяновский государственный технический университет, г. Ульяновск, 432027, Россия

Authors

Vladislav N. Kovalnogov [1], Dr.Sc. (Engineering), Head of Department "Thermal and Fuel Energy", kvn@ulstu.ru

Vyacheslav V. Sherkunov [1], Postgraduate Student, v.sherkunov@ulstu.ru

Hussein Mohamed Hussein [1], Postgraduate Student, mohammedab634@gmail.com

Vladimir N. Klyachkin [1], Dr.Sc. (Engineering), Professor of Department "Applied mathematics and informatics", v_kl@mail.ru

[1] Ulyanovsk State Technical University, Ulyanovsk, 432027, Russian Federation