УДК 004.9
DOI 10.25205/1818-7900-2020-18-4-11-27

Модель для прогнозирования температуры заготовки по ретроспекции ее нагрева на основе бустинга структуры «случайный лес»

П. И. Жуков, А. И. Глущенко, А. В. Фомин

Старооскольский технологический институт им. А. А. Угарова (филиал)
НИТУ «МИСиС»
Старый Оскол, Россия

Аннотация

Рассматривается проблема прогнозирования температуры поверхности стальной заготовки в прокатном стане после ее нагрева в методической печи. Эту оценку необходимо получить еще до того, как металл покинет печь. В отличие от классического подхода, основанного на решении краевой задачи теплопереноса на основе дифференциального уравнения нестационарной теплопроводности, в данном случае предлагается строить модель зависимости температуры заготовки от истории ее нагрева на основе анализа данных, полученных из системы управления печью. Собраны данные из АСУ ТП печей нагрева, и сформировано хранилище для них, проведен разведочный анализ данных, и определены объемы выборок для обучения, тестирования и валидации моделей. В рамках данной работы проведена валидация ранее разработанной авторами регрессионной модели. Ее результаты показали, что такой подход демонстрирует признаки переобучения (ошибка на проверочных выборках существенно превышает ошибку на обучающем множестве). Для того чтобы преодолеть указанный недостаток, в работе представлен альтернативный подход к построению искомой зависимости, основанный на поиске агрегированной гипотезы – бэггинга и бустинга. Результатом работы стало построение бустинг-модели «случайного леса» на основе особого класса классификационно-регрессионных деревьев – Dropout Adaptive Regression Trees (DART). На основе множественного эксперимента с полученной моделью были построены два доверительных интервала – 68 %-й и 95 %-й, а также рассчитано математическое ожидание ошибки прогноза ~ 9 °C по прогнозируемой температуре заготовки на стане как на обучающей, так и на валидационной выборке.

Ключевые слова

методические печи, прогнозирование температуры, разведочный анализ, технологические данные, бустинг, бэггинг, случайный лес, DART, MART

Для цитирования

Жуков П. И., *Глущенко А. И.*, *Фомин А. В.* Модель для прогнозирования температуры заготовки по ретроспекции ее нагрева на основе бустинга структуры «случайный лес» // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, № 4. С. 11–27. DOI 10.25205/1818-7900-2020-18-4-11-27

Prediction Model of Temperature of Cast Billet Based on Its Heating Retrospection Using Boosting “Random Forest” Structure

P. I. Zhukov, A. I. Glushchenko, A. V. Fomin

A. A. Ugarov Stary Oskol Technological Institute (Branch) NUST “MISIS”
Stary Oskol, Russian Federation

Abstract

The scope of this research is the prediction of a cast billet surface temperature, which it will have in the rolling mill after the heating process. The main problem is that such a prediction is needed before the cast billet will really leave

the furnace. In many cases, the boundary value problem of the heat transfer, particularly the differential equations of the transient heat conduction, is used to solve this problem. But in this research an alternative data-driven approach is proposed, which is based on a model of the dependence of the billet temperature on the retrospection of its heating in the continuous furnace. Such a model is developed as a result of the analysis of the data from the furnace control system. Such data from the real furnace were collected and stored in the data warehouse. Their exploratory analysis was conducted. All data were splitted into training, testing and validation subsets. As a part of this research, the regression model previously developed by the authors was also validated. It seemed to be overfitted (the error on the test set was significantly higher than the one on the training set). To overcome this disadvantage, an alternative method to develop the required data-based model is proposed by authors on the basis of the Boosting and Bagging algorithms. They belong to the machine learning field. As a result of the experiments with the bagging and boosting, the required model structure was chosen as a "Random Forest" with special class of the regression trees known as DART (Dropout Adaptive Regression Trees). Based on a significant number of experiments with that model, the two confidence intervals of the temperature prediction were found: 68 % and 95 % ones. The mean value of the temperature prediction error was estimated as ~ 9 °C for both the test and validation sets.

Keywords

continuous furnace, temperature prediction, exploration analysis, process data, boosting, bagging, random forest, DART, MART

For citation

Zhukov P. I., Glushchenko A. I., Fomin A. V. Prediction Model of Temperature of Cast Billet Based on Its Heating Retrospection Using Boosting "Random Forest" Structure. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 4, p. 11–27. (in Russ.) DOI 10.25205/1818-7900-2020-18-4-11-27

Введение

В течение последних десяти лет черная металлургия остается одним из крупнейших потребителей природных энергоносителей и вырабатываемой электроэнергии среди всех отраслей промышленности [1]. Структура энергопотребления на отечественных производствах не является исключением. Установлено, что разница в потреблении энергоресурсов наблюдается не только между странами, но и между отдельно взятыми предприятиями внутри одного государства [2]. Данный факт связан и с реализуемой технологией (доменные или электросталеплавильные печи), и с моральным устареванием действующих информационных систем, износом технологического оборудования. Данный факт приводит к потребности в оптимизации энергоемких технологических процессов не только на уровне самих технологических объектов, но также и на информационном уровне, проводя аналитическую модернизацию действующих систем управления.

На сегодняшний день в прокатных цехах металлургических производств имеют широкое распространение пламенные печи, в которых заготовки нагреваются в среде горения топлива перед тем, как будут переданы на прокатный стан. Для таких технологических объектов характерно большое потребление природного газа для поддержания высоких технологических температур (~ 1000–1200 °C). При этом нестационарность процесса нагрева (различные массы садок, скорость прокатки и, как следствие, скорость движения материала по печи, режимы нагрева) приводит к тому, что диапазон значений температуры заготовки на выходе из печи является достаточно широким даже в рамках одной технологической карты нагрева. В связи с этим формируется потребность знать температуру заготовки в момент ее захода на прокатный стан заранее (еще в тот момент, когда она находится в печи). Наличие подобных сведений позволит оператору более точно регулировать процесс нагрева, что в некоторых случаях поможет избежать нарушения технологии (перегрев или недогрев заготовки) и оптимально расходовать энергоресурсы.

В настоящее время поиск решения подобной задачи ведется в области аналитического моделирования, в частности решения краевых задач теплопроводности. Для методических печей нагрева перспективным является математическое моделирование на основе сеточных моделей аппроксимации дифференциального уравнения нестационарной теплопроводности, где граничные условия задачи описывают различный контекст нагрева [3; 4]:

$$\rho * c\frac{\partial t}{\partial \tau} = \mathrm{div}\left(\lambda * \mathrm{grad}\left(t\right)\right).$$

Здесь ρ – плотность металла, кг/м^3; c – теплоемкость металла, Дж / (кг*К); λ – теплопроводность металла, Вт/(м*К).

Основная проблема данного подхода заключается в противоречивых требованиях, которые предъявляются к конечным моделям, а именно: 1) высокая точность; 2) возможность расчета в режиме реального времени. В настоящее время достаточно точное численное решение (ровно, как и аналитическое) краевых задач теплопроводности в полной постановке не всегда существует, и исходную задачу приходится упрощать, принимая определенного рода допущения, что негативно сказывается на конечной точности модели. При этом граничные условия (например, геометрические), конструктивные особенности печей и контекст нагрева требуют адаптации подобных упрощенных моделей [5] с целью повышения их конечной точности, при этом делая их менее универсальными.

С другой стороны, в настоящий момент вместе с развитием информационных технологий и ростом мощности ЭВМ получили развитие не только численные методы решения дифференциальных уравнений, но и интеллектуальные подходы и методы анализа данных [6; 7]. Одновременно с этим современные АСУ ТП в процессе функционирования собирают, обрабатывают и сохраняют большие объемы информации, которые, предположительно, наполнены полезными для решения рассматриваемой задачи зависимостями. Предполагается, что такая информация может отражать не только опыт и навыки экспертов, взаимодействующих с технологическим объектом, но также и теплофизические процессы в виде линейных и нелинейных зависимостей в данных.

Опираясь на вышеизложенные факты, в данной работе было проведено исследование информации, полученной из АСУ ТП печей нагрева, с целью построения модели зависимости температуры заготовки на стане от времени ее нагрева и температуры в зонах печи. Исходными технологическими объектами управления являлись печи нагрева № 1 и 2 сортопрокатного цеха № 1 (СПЦ-1) АО «Оскольский электрометаллургический комбинат».

Сбор данных и проектирование хранилища

Объектом исследования являлась шестизонная пламенная печь нагрева с шагающим механизмом и попарным распределением зон (четные зоны – под, нечетные – свод). В качестве исходных данных авторами была получена информация из действующих подсистем АСУ ТП печи, эквивалентная 62-м дням работы объекта.

Полученные данные имели вид трех обособленных наборов. В результате анализа их структуры были выявлены основные опорные точки (ключевые параметры), по которым можно осуществить связь имеющихся данных. Для того чтобы хранить результирующий набор, полученный после первичных преобразований, было решено спроектировать хранилище. Для этого нужно представить все имеющиеся данные в виде классической OLAP-фигуры:

$$D:\left(X_1, X_2, ..., X_N\right) \to W_k.$$

Здесь D – это набор данных, X_1, X_2, X_N – атрибуты данных, которые формируют измерения (оси), N – количество эти параметров, W_k – OLAP-фигура, содержащая меры (значения) по всем измерениям (осям) следующей фигуры:

$$W_k = \left\{\overrightarrow{w_{X_1}}, \overrightarrow{w_{X_2}}, ..., \overrightarrow{w_{X_N}}\right\}.$$

Проекция W_k по любому значению $k \in [1, 2, \ldots, N]$ будет представлять собой кортеж значений, а следовательно, для физической реализации целесообразнее будет использовать

структуру, наиболее близкую к реляционным базам данных. Таким образом, в качестве виртуальной структуры были выбраны двумерные матрицы (таблицы вида «dataframe» в языке R), а для физического хранилища – файловая система.

Хранилище, реализованное реляционной схемой «звезда», изображено на рис. 1. Здесь параметры «Номер плавки», «Номер проката» и «Единица проката» являются вспомогательными атрибутами, которые помогают алгоритмам предобработки однозначно идентифицировать заготовку; параметры FTZ, STZ и TTZ отвечают за время, которое заготовка провела в 1, 2 и 3-й паре зон соответственно. Аналогичным образом параметры с префиксом *T*1, *T*2 и *T*3 отвечают за соответствующие температуры в 1, 2, и 3-й паре зон соответственно в моменты, когда в них находилась заготовка (причем температура хранится в виде диапазона от минимальной *Ti*Min до максимальной *Ti*Max, $i = 1, 2, 3$); параметр *Trez* отвечает за температуру заготовки, снятую пирометром в прокатном стане. Таким образом, для дальнейшего исследования был сформирован набор данных из 9 независимых переменных и одной зависимой переменной (*Trez*).

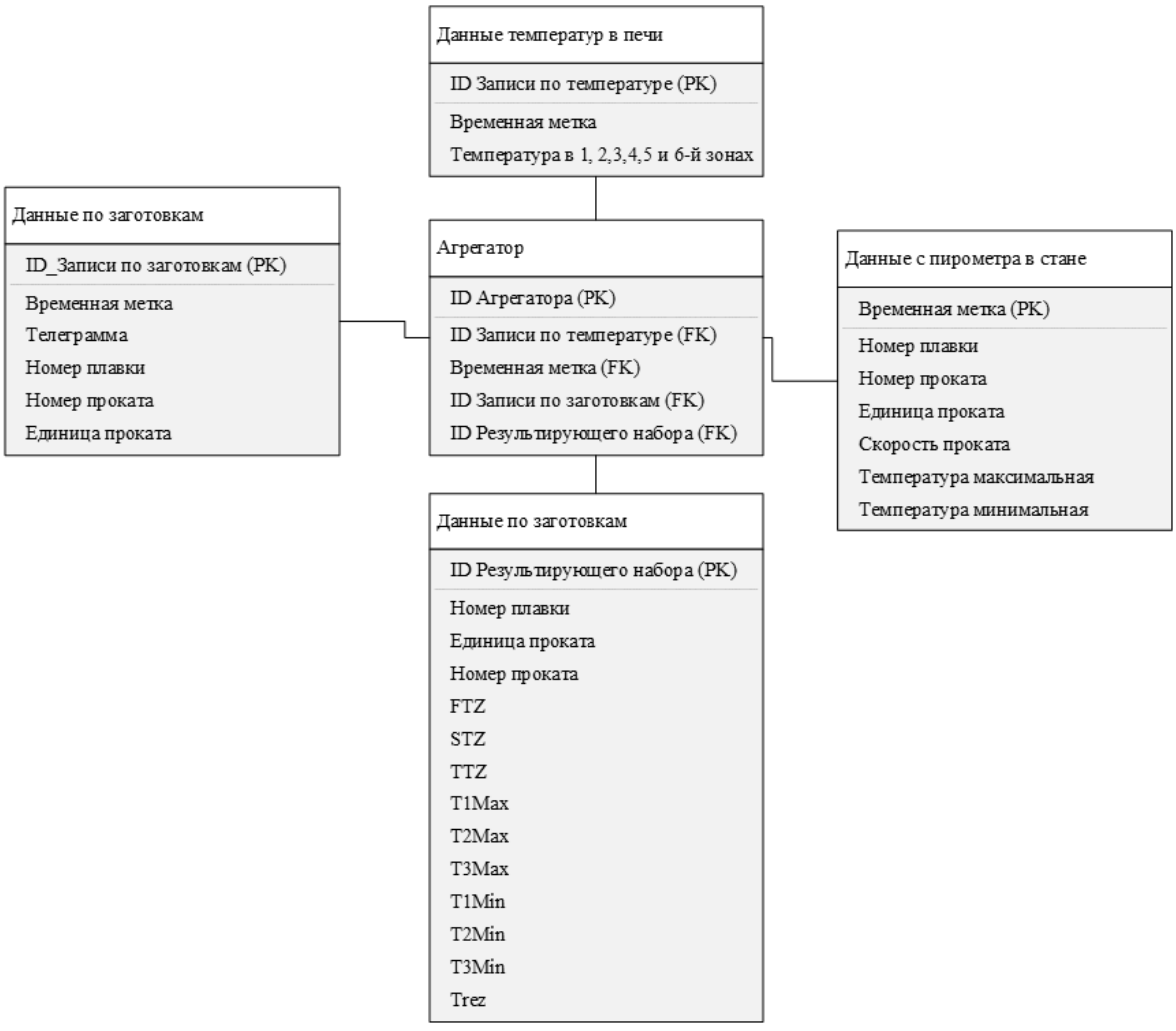

Рис. 1. Хранилище данных (Статичная структура в нотации UML)
Fig. 1. Data Warehouse using UML notation

В начале исследования хранилище удалось наполнить 7556-ю записями. Данные были помечены как экспериментальные (первичные) для построения искомой зависимости. Далее получены дополнительные данные за 31-й день работы печи и обработаны в 1920 записей, которые были помечены как данные для валидации найденных моделей (валидационные).

Разведочный анализ данных

В качестве исходной точки анализа было решено начать с определения точек-выбросов, так как робастность большинства методов зависит от данного параметра [8]. Поэтому необходимо либо обнаружить и устранить такие наблюдения, либо пересмотреть применяемые методы. Для определения выбросов использовалась классическая диаграмма размахов с медианной оценкой и полуторным интерквартильным расстоянием, пример которой представлен на рис. 2. Здесь были рассмотрены параметры из первичного набора данных. Для остальных параметров в рамках этого набора аналогичным образом наблюдается большое количество выбросов. На данном этапе выдвинуто предположение о том, что количество выбросов характерно для вероятностного распределения неизвестного рода, отличного от нормального, и не является следствием ошибки в сборе данных.

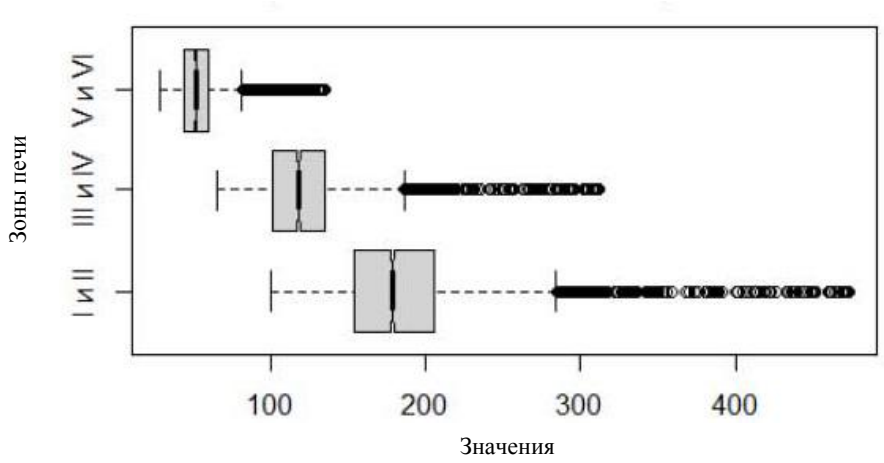

Рис. 2. Пример распределения выбросов в первичных данных на основе диаграммы размахов
Fig. 2. Outliers in Cast billet time parameters (primary data)

Для проверки выдвинутого предположения были проведены формальные тесты. Основой для формального теста на нормальность послужил критерий Колмогорова – Смирнова, который используется для проверки нулевой гипотезы на соответствие случайной величины X некоторому известному закону распределения $F(x)$ и имеет статистику вида

$$D_n = \sup\left|F_n(x) - F(x)\right|, \quad (1)$$

где $F_n(x)$ – некоторая эмпирическая функция распределения; $F(x)$ – некоторая функция распределения с известными параметрами. Для практической проверки соответствия независимых переменных выборки нормальному распределению было решено использовать критерий (1) в интерпретации Лиллиефорса:

$$D_n^* = D_n(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}).$$

Здесь n – объем выборки. Выбор данного критерия обусловливается его статистической мощностью, а также возможностью применить его к выборкам с большим значением n [9].

Для нахождения контрольного значения рассчитанных критических статистик использовалась формула

$$\lambda_{\alpha,N} = \frac{\lambda_{\alpha,0}}{\sqrt{N}}, \quad (2)$$

где $\lambda_{\alpha,N}$ – табличное (контрольное) значение статистики для уровня значимости α при объеме выборки N; $\lambda_{\alpha,0}$ – контрольный коэффициент для заданного уровня значимости α. Таким образом, при уровне значимости 0.05 коэффициент $\lambda_{\alpha,0}$ будет равен 0.886. Исходя из объема выборки 7556 записей, контрольное значение статистики равно 0.01019266.

Результаты формального теста для некоторых независимых переменных представлены в табл. 1.

Таблица 1

Результаты теста Лиллиефорса

Table 1

Lilliefors test results

Переменные	$D_n^*(x)$	p-значение
FTZ = STZ = TTZ	0.091214	$< 2.2*10^{-16}$
*T*1Max	0.24897	$< 2.2*10^{-16}$
*T*2Max	0.25398	$< 2.2*10^{-16}$
*T*3Max	0.16835	$< 2.2*10^{-16}$
Trez	0.068631	$< 2.2*10^{-16}$

В результате формального теста было установлено, что среди рассмотренных параметров наиболее близки к критическому значению параметры, связанные со временем нахождения заготовки в каждой паре зон, а также целевой критерий. Для параметров температур отсутствует нормальное распределение, так как рассчитанные статистики на порядок превышают контрольное значение. Для полноты анализа было решено построить также гистограммы распределения, но используя в качестве оценочного критерия плотность, а не частоту. Такой подход позволит совместить гистограмму с графиком плотности.

Основываясь на рис. 3, было сделано предположение, что вероятностное распределение данного параметра (FTZ), а также параметров, однотипных ему (STZ, TTZ), принадлежит семейству вероятностей Пирсона IX со смещенным математическим ожиданием влево относительно центра интервала.

Рассмотрев также и зависимую переменную, был сделан аналогичный вывод, и принято допущение о нормальности распределения следующих параметров: *FTZ*, *STZ*, *TTZ* и *Trez*. Для остальных параметров нормальность распределения не наблюдается.

Помимо рассмотренных выше тестов, были также проведены следующие этапы разведочного анализа: тест на линейные зависимости среди переменных; корреляционные тесты; тесты на избыточные нули и пропущенные значения [8]. В результате анализа было установлено, что среди данных имеются слабые значимые корреляции независимых отдельных наблюдений и зависимой переменной, а также отсутствуют коллинеарные зависимости.

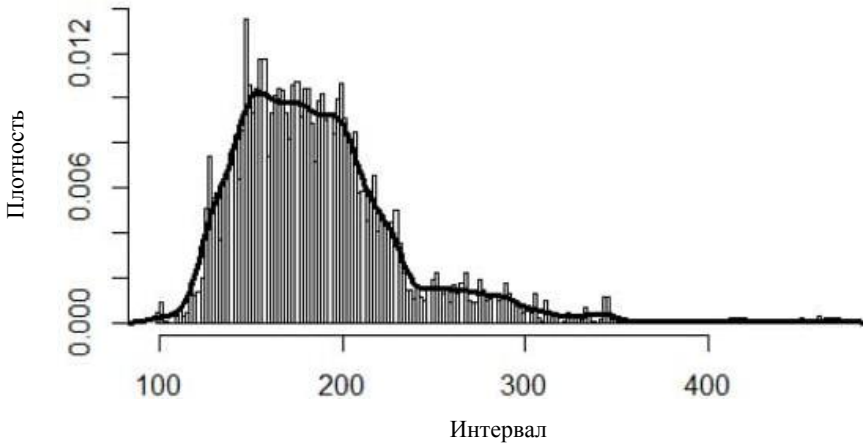

Рис. 3. Пример гистограммы и функции плотности
для параметра «Время заготовки в первой паре зон»
Fig. 3. Visual analysis of distribution density
of "Cast Billet time in 1st pair of zones" regressor

Перед проведением дальнейших исследований было решено разбить имеющиеся данные следующим образом (табл. 2).

Таблица 2

Разбиение исходных данных

Table 2

Data subsampling

Первичные данные (7556 записей), в том числе			Валидационные данные (1920 записей), в том числе	
для обучения	для тестирования	для валидации	для валидации	контрольное множество
6047	1509	–	1537	383

Разбиение данных происходило с сохранением второго момента целевой переменной

$$\mu_2^1 = \mu_2^2 = \mu_2^3 = ... = \mu_2^k, \quad \forall k \in \mathbb{N}, \tag{3}$$

где μ_2^k – дисперсия *k*-го подмножества исходного множества данных.

Тестирование регрессионной модели

Ранее было проведено исследование применимости планов регрессионного анализа для построения искомой зависимости [10] при допущении о нормальности ключевых переменных (регрессоров). В результате такого исследования была получена модель на основе следующего регрессионного уравнения:

$$f\left(y_j\right) = \beta_0 + \sum_{i=1}^{6}\sum_{k=1}^{p}\beta_{ik} * x_{ij}^k + \sum_{m=1}^{4}\sum_{k=1}^{p}\beta_{mk} * \left(x_{1j}^m * x_{3j}^{(p+1)-k}\right). \tag{4}$$

Здесь $f(y_j)$ –значение температуры заготовки на стане после нагрева в печи; $p = 4$ – степень используемого полинома; $m = 4$ – коэффициент, позволяющий учесть необходимое количество комбинаторных взаимодействий для выбранных переменных; x_{1j} – j-е значение параметра «FTZ» (см. рис. 1); x_{3j} – j-е значение параметра «TTZ» (см. рис. 1).

На первом этапе было решено протестировать полученную модель (4) на данных для валидации (см. табл. 2), к которым не было доступа на момент проведения исследования [10]. Результаты тестирования представлены на рис. 4.

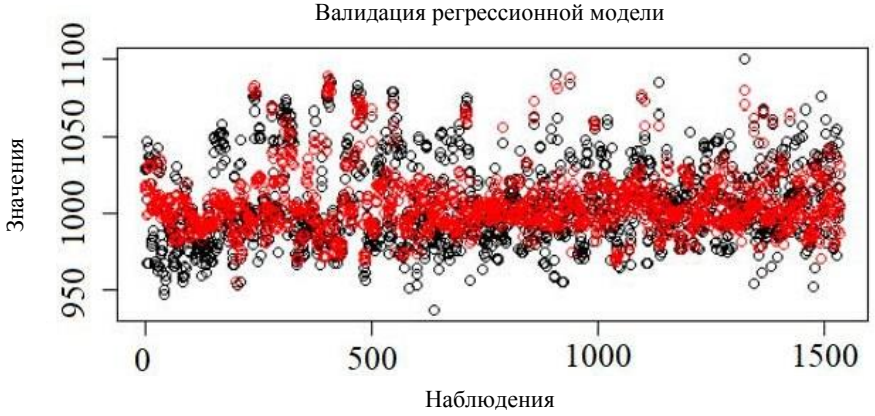

Рис. 4. Результаты валидации регрессионной модели (4) на валидационных данных (красные точки – прогноз модели; черные точки – исходные данные)
Fig. 4. Results of regression model validation (red points are model outputs; black points are ground truth)

Валидация модели показала, что наблюдается рост средней ошибки прогноза модели на новых данных, не задействованных ранее в процессе исследований, с 15 °C в исследовании [10] на тестовом множестве до 25 °C на валидационном множестве, при сохранении средней температуры 1030 °C. Данный факт может свидетельствовать как о слабой устойчивости полученной модели, так и о ее переобучении.

Дальнейшие эксперименты с обобщенными моделями не позволили добиться существенного увеличения точности прогноза. Исходя из этого было принято решение отойти от попыток построить одну комбинированную модель и искать решения среди слабых моделей, агрегированных между собой.

Построение композиционной модели

На сегодняшний день наибольшую популярность среди планов поиска агрегированной гипотезы получили композиционные алгоритмы: бэггинг и бустинг [11]. Данные подходы позволяют комбинировать слабые, с точки зрения обобщения, модели в сильный решающий ансамбль путем усиливающего объединения и усиливающего пересечения соответственно [12–15]. При работе с бустинг- и бэггинг-структурами было принято решение использовать классификационно-регрессионные деревья (далее деревья решений).

В процессе экспериментов с композиционными структурами, наилучшие результаты были получены при использовании библиотеки экстремального градиентного бустинга (XGBoost) [16]. Основной идеей композиционной модели было совмещение бэггинг-структур вида «случайный лес» с алгоритмом экстремального градиентного бустинга.

После построение дерева решений (бэггинг-структуры), выход дерева подается на общий оптимизатор следующего вида:

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i - \hat{y}_i^{(t-1)},\ f_t\left(x_i\right)\right) + \Omega\left(f_t\right).$$

Здесь l – двумерная функция потерь от невязки $y_i - \hat{y}_i^{(t-1)}$ (разницы между значением i-го элемента выборки и суммой предсказаний первых t деревьев) и выхода t-го дерева $f_t(x_i)$, обученного на подмножестве x_i из множества входов X и отобранного алгоритмом формирования дерева в случайном порядке с сохранением нормального вероятностного распределения случайной величины; x_i – набор независимых переменных (признаков) для i-го элемента выборки; $\Omega(f_t)$ – это регуляризация t-го дерева решений f_t, N – количество наблюдений результирующей переменной.

XGBoost оперирует понятием «дерево» в своей интерпретации: подразумевается иерархическая структура, в листах которой может находиться модель совершенно любого вида. Применимо к деревьям решений данный алгоритм представляется в следующем виде (рис. 5).

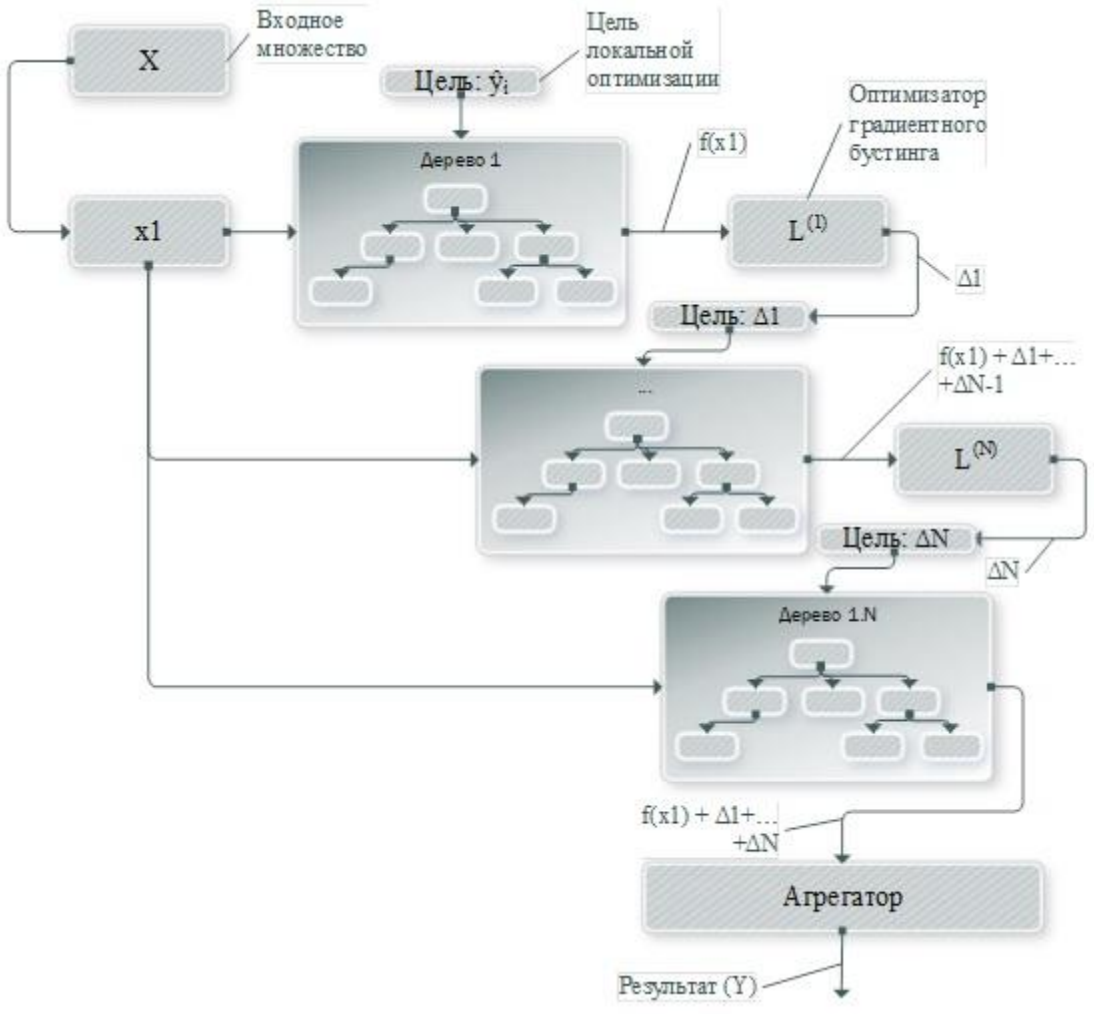

Рис. 5. Функциональная схема градиентного бустинга деревьев решений
Fig. 5. Functional diagram of gradient boosting of decision trees

Для практической реализации данной структуры были использованы язык R и соответствующая библиотека. Параметризация алгоритма, подобранная экспериментально, представлена в табл. 3.

Таблица 3

Параметризация модели

Table 3

Model parameterization

Параметр	Значение
Количество раундов обучения	200
Количество параллельно обучаемых деревьев	50
Максимальная глубина 1-го дерева	10
Процент данных, видимый 1-м деревом	0.8 (80 % данных для 1-го дерева)
Процент данных, видимый всей моделью	1 (100 % данных)

Средняя ошибка прогноза на тестовых данных (см. табл. 2) составила 11 °C по прогнозируемой температуре заготовки на стане (рис. 6). Данный показатель точности вдвое превышает результаты валидации модели (2). Точки аппроксимации больше не сконцентрированы вокруг математического ожидания и с хорошей плотностью покрывают значения на концах интервала, а тенденция сохраняется на всем интервале наблюдений.

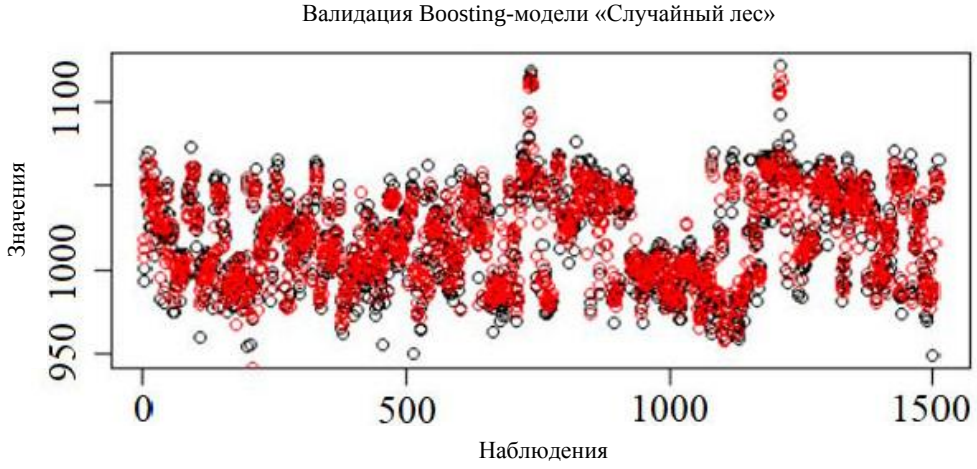

Рис. 6. Результаты валидации Boosting-модели «случайный лес» на тестовых данных
Fig. 6. Result of validation of boosting random forest model
(red points are model prediction; black points are ground truth values)

Однако апробация модели на данных для валидации показала систематичное падение точности как для отдельно взятых деревьев решений, так и для всей модели в целом (табл. 4).

При этом было установлено, что дерево решений на валидационном множестве теряет больше по точности в абсолютных числах (см. табл. 4), что свидетельствует о наличии переобучения вида «сверхспециализация». Данный тип переобучения характеризуется концентрацией модели на точках-выбросах и попытках «подгонки» функции ошибки под краевые значения интервала, что приводит к систематичному падению точности. Для проверки всей

модели было решено оценить точность на контрольных данных (см. табл. 2). В результате было установлено, что рост ошибки наблюдается с 11 до 17 °C в абсолютных значениях при средней температуре заготовки 1030 °C.

Таблица 4

Результаты валидации полученной модели

Table 4

Validation results

Модель	Средняя ошибка прогноза в абсолютных значениях температуры, °C	
	тестовые данные	валидационные данные
Отдельно взятое дерево	12	16
Boosting-модель	11	13

Было выдвинуто предположение, что подобное переобучение вызвано недостаточным количеством имеющихся регрессоров, что в итоге косвенно подтверждается несоответствием построенной модели контексту предметной области (рис. 7), а следовательно, и невозможности с высокой точностью аппроксимировать данные с реальных (действующих) систем слежения за металлом.

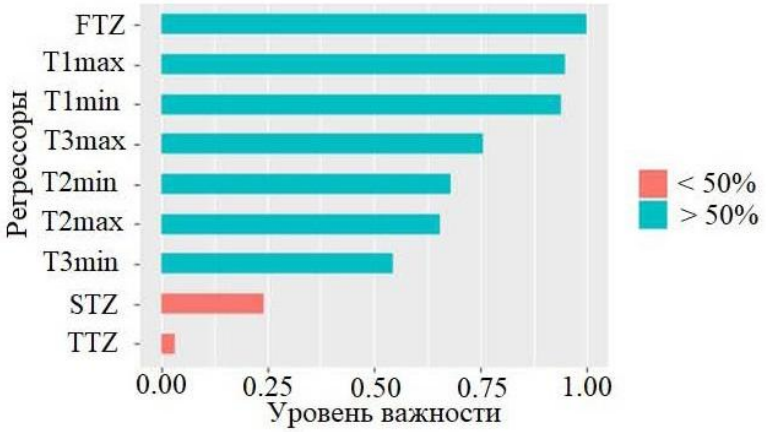

Рис. 7. Уровни важности регрессоров для полученной модели
Fig. 7. Importance level of model regressors

Было установлено, что модель выделяется, среди прочего, важными параметры *T*1*max* и *T*1*min*, в то время как по условиям реального производства в данной зоне используются более низкие температуры по сравнению с другими зонами печи. На основании имеющихся результатов был сделан вывод о необходимости качественного расширения имеющихся данных.

Для дальнейших экспериментов было решено увеличить количество независимых переменных. За основу было взято хранилище для валидации, и оно было расширено (рис. 8). Обновленный набор данных по-прежнему содержал 1920 записей, но качественно увеличилось количество объясняющих переменных (с 9-ти до 23-х). Полученное множество было разбито на: обучающее (1442 значения), тестовое (428 значений) и валидационное (50 значе-

ний) подмножества по аналогии с предыдущим набором данных, сохраняя вторые моменты получившихся подмножеств (3).

Первичные (7556 записей) данные более не использовались при проведении экспериментов. Для них расширение регрессоров было невозможно. Кроме того, они уже позволили определить необходимый подход к построению требуемой зависимости – бустинг-структуры «случайный лес».

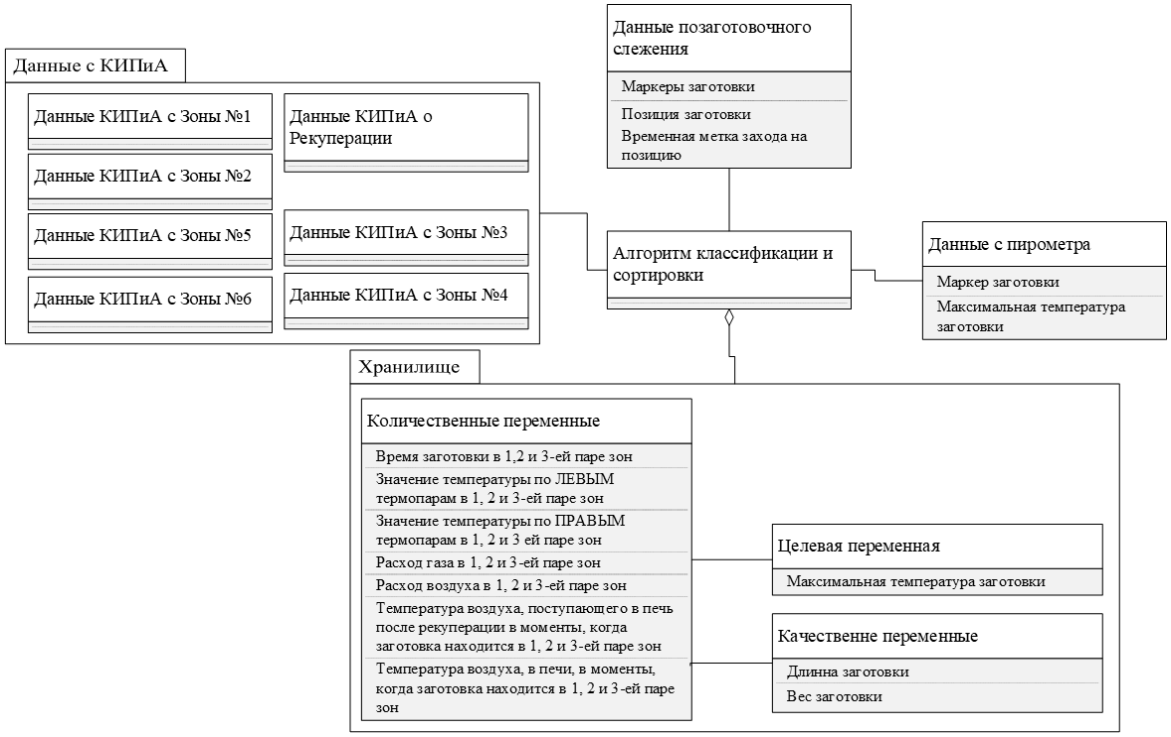

Рис. 8. Обновленное хранилище данных в статичной структуре нотации UML
Fig. 8. Extended data warehouse using UML notation

Для того чтобы в дальнейшем избежать ситуации «сверхспециализации», был задействован модифицированный алгоритм градиентного бустинга – DART, в основе которого лежит технология Dropout, хорошо зарекомендовавшая себя в алгоритмах глубокого обучения искусственных нейронных сетей [17]. Параметризация алгоритма, также определенная экспериментальным путем, представлена в табл. 5.

При этом ошибку модели на тестовых данных удалось снизить до 9 °С по прогнозируемой температуре заготовки на стане. Множественный эксперимент (повторное разбиение данных с другим перцентилем на описанные выше три подмножества) и апробация на получавшихся в результате валидационных данных показали, что наблюдается устойчивый интервал получаемой ошибки: от 9 до 11° по прогнозируемой температуре заготовки на стане при средней температуре заготовки 1000 °С.

При этом выборка из результатов множественного эксперимента (среднеквадратичная ошибка на один раунд обучения) имеет нормальное вероятностное распределение. Возникающий в процессе апробации устойчивый интервал, а также факт нормального распределения ошибки позволяет с высокой доверительной вероятностью перейти от точечной оценки к интервальной с допущением о минимальной погрешности.

Таблица 5

Параметризация DART-модели

Table 5

DART-model parameterization

Параметр	Значение
Количество раундов обучения	400
Коэффициент Dropout	0.5
Количество параллельных деревьев	25
Максимальная глубина дерева	10
Количество данных, видимых в 1-м раунде обучения	0.25 (25 % данных на раунд)
Количество данных, видимых единовременно одному дереву в модели	0.6 (60 % данных на дерево с повторениями)

Основываясь на вышеизложенном, было решено формировать итоговую оценку в виде доверительных интервалов, основываясь на «правиле трех сигм» нормального распределения. Результаты тестирования полученной модели в интервальном виде представлены на рис. 9.

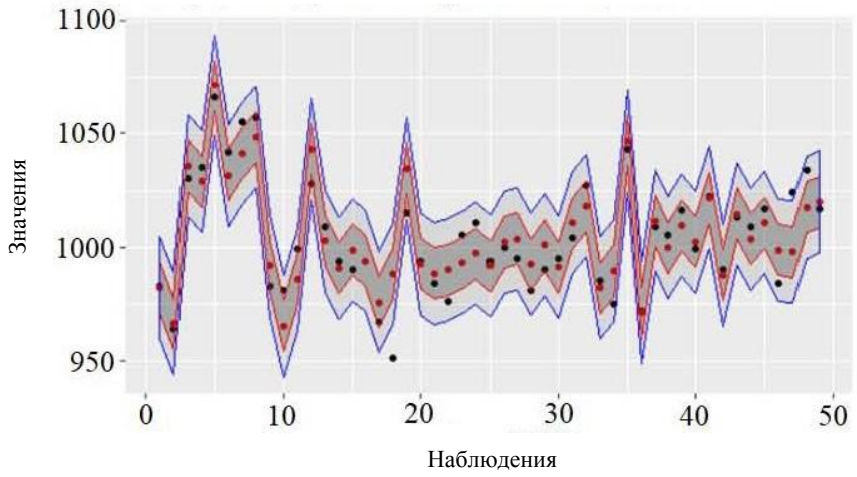

Рис. 9. Результаты валидации модели DART-бустинга с интервальной оценкой
(красные точки – прогнозы модели; черные точки – валидационные данные;
внутренняя область – 68 % доверительный интервал;
внешняя область – 95 % доверительный интервал)
Fig. 9. Result of DART model testing
(red points are model predictions; black points are test values;
red lines form 68 % confidence interval; blue lines form 95 % confidence interval)

Заключение

В процессе исследования выделены характерные особенности технологических данных: большое количество выбросов, отсутствие нормальных распределений температурных параметров (регрессоров). На основе данного наблюдения сделан вывод о нецелесообразности применения обобщенных регрессионных моделей для построения искомой зависимости. Решение найдено среди методов машинного обучения, и в результате была получена DART-модель градиентного бустинга «Случайного леса» с интервальной оценкой точности:

$\hat{Y} \pm 11$ °C для 68 % доверительного интервала и $\hat{Y} \pm 22$ °C – для 95 % доверительного интервала, где $\hat{Y}$ – множество прогнозов модели (т. е. прогнозируемая температура заготовки на стане). Математическое ожидание ошибки прогноза составляет $M\left(Y - \hat{Y}\right) = 9$ °C. Полученная модель способна функционировать в реальном времени.

Данной точности достаточно для применения подобной системы в качестве информирующей надстройки для операторов и технологов, взаимодействующих с объектом.

В дальнейшем авторами планируется увеличить количество регрессоров, а также увеличить объем выборки. Предполагается, что увеличение повторяемости в данных позволит исследовать поведение точечной оценки прогноза в контексте различных рабочих режимов исследуемого объекта, а также сузить интервал ошибки прогноза. В процессе оценки полученных результатов было выдвинуто предположение, что существует достаточно узкий интервал ошибки прогноза, при котором данную модель можно было бы задействовать в контуре управления объектом.

Список литературы

1 **Рудыка В. И.**, **Малина В. П**. Сталь, кокс, уголь в 2010 г. и далее – состояние, посткризисные прогнозы и перспективы // Кокс и химия. 2010. № 2. С. 2–11. DOI 10.3103/s1068364x1012001x

2 **Новиков Н. И., Новикова Г. В**. Топливно-энергетическая составляющая черной металлургии: проблемы и тенденции // Вестник КемГУ. 2013. № 4 (56). С. 257–263.

3 **Бирюков А. Б., Волошин А. И., Олешкевич Т. Г**. Математическое моделирование процесса тепловой обработки металла в печах // Сталь. 2016. № 1. С. 71–75.

4 **Бирюков А. Б., Гинкул С. И., Гнитиев П. А., Олешкевич Т. Г**. Математическое моделирование процессов тепловой обработки металла в печах с учетом окалинообразования // Сталь. 2016. № 8. С. 85–90.

5 **Бирюков А. Б., Гнитиёв П. А., Олешкевич Т. Г**. Адаптация математической модели процессов тепловой обработки металла в печах, учитывающей окалинообразование // Вестник Донецкого нац. техн. ун-та. 2017. № 2 (8). С. 30–37.

6 **Саранча С. Ю., Моллер А. Б**. Применение информационных технологий в металлургическом производстве: оптимизация технологии прокатки и раскроя готовой продукции в сортопрокатном производстве // Актуальные проблемы современной науки, техники и образования. 2014. Т. 1. С. 139–143.

7 **Беренов Д. А., Белан С. Б., Аксенов К. А., Перескоков С. А**. Полностью оцифрованное металлургическое производство: слежение, аналитика, моделирование // Фундаментальные исследования. 2017. № 9-2. С. 272–277.

8 **Zuur, Alain F., Elena N. Ieno, Chris S. Elphick**. A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, 2010, vol. 1, no. 1, p. 3–14. DOI 10.1111/j.2041-210x.2009.00001.x

9 **Zacharias P., Vávra M**. A distance test of normality for a wide class of stationary processes. *Econometrics and Statistics*, 2017, vol. 2, p. 50–60. DOI 10.1016/j.ecosta.2016.11.005

10 **Жуков П. И., Глущенко А. И., Фомин А. В**. Построение зависимости температуры непрерывно литой заготовки от ретроспекции её нагрева // Системы управления и информационные технологии. 2019. № 4 (78). С. 73–78.

11 **Zhou Z**. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 2013, vol. 203, p. 1–18. DOI 10.1016/j.artint.2013.07.002

12 **Basha, Syed Muzamil, Dharmendra Singh Rajput, Vishnu Vandhan**. Impact of gradient ascent and boosting algorithm in classification. *International Journal of Intelligent Engineering and Systems* (*IJIES*), 2018, vol. 11 no. 1, p. 41–49. DOI 10.22266/ijies2018.0228.05

13 **Gomes, Heitor M. et al**. Adaptive random forests for evolving data stream classification. *Machine Learning*, 2017, vol. 106, no. 9–10, p. 1469–1495. DOI 10.1007/s10994-017-5642-8
14 **Khiari, Jihed. et al**. Metabags: Bagged meta-decision trees for regression. Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, 2018. DOI 10.1007/978-3-030-10925-7_39
15 **Döpke J., Fritsche U., Pierdzioch C**. Predicting recessions with boosted regression trees. *International Journal of Forecasting*, 2017, vol. 33 no. 4, p. 745–759. DOI 10.1016/j.ijforecast.2017.02.003
16 **Qian, Ning et al.** Predicting heat transfer of oscillating heat pipes for machining processes based on extreme gradient boosting algorithm. *Applied Thermal Engineering*, 2020, vol. 164. DOI 10.1016/j.applthermaleng.2019.114521
17 **Vinayak R. K., Gilad-Bachrach R**. Dart: Dropouts meet multiple additive regression trees. *Artificial Intelligence and Statistics, PMLR*, 2015, vol. 38, p. 489–497.

References

1. **Rudyka V. I., Malina V. P**. Steel, coke, and coal in 2010 and beyond: Prospects beyond the crisis. *Coke Chem*., 2010, vol. 53, p. 433–442. DOI 10.3103/S1068364X1012001X
2. **Novikov N. I., Novikova G. V**. Fuel and energy component of ferrous metallurgy: problems and tendencies. *Vestnik KemSU*, 2013, vol. 4, no. 56, p. 257–263. (in Russ.)
3. **Biryukov A. B., Voloshin A. I., Oleshkevich T. G**. Matematicheskoe modelirovanie protsessa teplovoi obrabotki metalla v pechakh [Mathematical simulation of the processes of thermal treatment of metal in furnaces]. *Stal'*, 2016, vol. 1. p. 71–75. (in Russ.)
4. **Biryukov A. B., Ginkul S. I., Gnitiev P. A., Oleshkevich T. G**. Matematicheskoe modelirovanie protsessov teplovoi obrabotki metalla v pechakh s uchetom okalinoobrazovaniya [Mathematical simulation of the processes of thermal treatment of metal in furnaces taking into account formation of scale]. *Stal'*, 2016, vol. 8, p. 85–90. (in Russ.)
5. **Biryukov A. B., Gnitiev P. A., Oleshkevich T. G**. Adaptation of the mathematical model of metal heat treatment in furnaces considering scale formation. *Vestnik DonNTU*, 2017, vol. 2, no. 8, p. 30–37. (in Russ.)
6. **Sarancha S. Yu., Moller A. B**. Primenenie informatsionnykh tekhnologii v metallurgicheskom proizvodstve: optimizatsiya tekhnologii prokatki i raskroya gotovoi produktsii v sortoprokatnom proizvodstve [Application of information technologies in metallurgical production: optimization of rolling technology and cutting of finished products in shape and bar production]. *Aktual'nye problemy sovremennoi nauki, tekhniki i obrazovaniya*, 2014, vol. 1. p. 139–143. (in Russ.)
7. **Berenov D. A., Belan S. B., Aksenov K. A., Pereskokov S. A**. Digitally of metallurgical production: monitoring, analytic, modeling. *Fundamental research*, 2017, vol. 9-2, p. 272–277. (in Russian)
8. **Zuur, Alain F., Elena N. Ieno, Chris S. Elphick**. A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, 2010, vol. 1, no. 1, p. 3–14. DOI 10.1111/j.2041-210x.2009.00001.x
9. **Zacharias P., Vávra M**. A distance test of normality for a wide class of stationary processes. *Econometrics and Statistics*, 2017, vol. 2, p. 50–60. DOI 10.1016/j.ecosta.2016.11.005
10. **Zhukov P.I., Glushchenko A.I., Fomin A.V**. Development of dependence of continuously cast billets temperature from its heating retrospection. *Sistemy upravleniya i informatsionnye tekhnologii*, vol. 4, no. 78, p. 73–78. (in Russ.)
11. **Zhou Z**. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 2013, vol. 203, p. 1–18. DOI 10.1016/j.artint.2013.07.002

12. **Basha, Syed Muzamil, Dharmendra Singh Rajput, Vishnu Vandhan**. Impact of gradient ascent and boosting algorithm in classification. *International Journal of Intelligent Engineering and Systems* (*IJIES*), 2018, vol. 11 no. 1, p. 41–49. DOI 10.22266/ijies2018.0228.05
13. **Gomes, Heitor M. et al**. Adaptive random forests for evolving data stream classification. *Machine Learning*, 2017, vol. 106, no. 9–10, p. 1469–1495. DOI 10.1007/s10994-017-5642-8
14. **Khiari, Jihed. et al**. Metabags: Bagged meta-decision trees for regression. Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, 2018. DOI 10.1007/978-3-030-10925-7_39
15. **Döpke J., Fritsche U., Pierdzioch C**. Predicting recessions with boosted regression trees. *International Journal of Forecasting*, 2017, vol. 33 no. 4, p. 745–759. DOI 10.1016/j.ijforecast.2017.02.003
16. **Qian, Ning et al.** Predicting heat transfer of oscillating heat pipes for machining processes based on extreme gradient boosting algorithm. *Applied Thermal Engineering*, 2020, vol. 164. DOI 10.1016/j.applthermaleng.2019.114521
17. **Vinayak R. K., Gilad-Bachrach R**. Dart: Dropouts meet multiple additive regression trees. *Artificial Intelligence and Statistics*, *PMLR*, 2015, vol. 38, p. 489–497.

Материал поступил в редколлегию
Received
09.09.2020

Сведения об авторах

Жуков Пётр Игоревич, аспирант, СТИ НИТУ «МИСиС», кафедра автоматизированных и информационных систем управления (Старый Оскол, Россия)

Zhukov.Petr86@yandex.ru
ORCID 0000-0002-8859-6739

Глущенко Антон Игоревич, кандидат технических наук, доцент, СТИ НИТУ «МИСиС», кафедра автоматизированных и информационных систем управления (Старый Оскол, Россия)

strondutt@mail.ru
ORCID 0000-0002-6948-9807

Фомин Андрей Вячеславович, кандидат технических наук, старший преподаватель, СТИ НИТУ «МИСиС», кафедра автоматизированных и информационных систем управления (Старый Оскол, Россия)

verner444@yandex.ru
ORCID 0000-0001-9867-2195

Information about Authors

Petr I. Zhukov, postgraduate student, A. A. Ugarov Stary Oskol Technological Institute (Branch) NUST “MISIS”, Automation and information control systems department (Stary Oskol, Russian Federation)

Zhukov.Petr86@yandex.ru
ORCID 0000-0002-8859-6739

Anton I. Glushchenko, Candidate of Sciences, Associate Professor, A. A. Ugarov Stary Oskol Technological Institute (Branch) NUST "MISIS", Automation and information control systems department (Stary Oskol, Russian Federation)

strondutt@mail.ru
ORCID 0000-0002-6948-9807

Andrey V. Fomin, Candidate of Sciences, Senior lecturer, A. A. Ugarov Stary Oskol Technological Institute (Branch) NUST "MISIS", Automation and information control systems department (Stary Oskol, Russian Federation)

Zhukov.Petr86@yandex.ru
ORCID 0000-0002-8859-6739