

Научная статья

УДК 004.051

URL: <https://trudymai.ru/published.php?ID=182676>

МОДЕЛИРОВАНИЕ ВЫБОРА ХАРАКТЕРИСТИК ВРЕМЕННЫХ РЯДОВ ДЛЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ КОРРЕЛЯЦИОННОЙ МАТРИЦЫ

**Павел Юрьевич Гусев^{1✉}, Александр Валентинович Таволжанский², Ирина
Валерьевна Поцебнева³**

^{1,2,3}Воронежский государственный технический университет,

Воронеж, Россия

¹pgusev@cchgeu.ru✉

²tavolzhanskij.a@yandex.ru

Аннотация. В данной работе рассматривается проблема выбора необходимого набора характеристик для прогнозирования временного ряда. Предлагается метод решения этой задачи на основе корреляционной матрицы. По подготовленным данным строится корреляционная матрица, после чего для каждой переменной формируется список, упорядоченный по убыванию модуля степени корреляции. Затем происходит обучение моделей линейной регрессии и сравнение качества прогнозов для различных наборов переменных из отсортированного списка. Далее осуществляется сравнение для различных значений времени дискретизации, для определения оптимального значения для каждой переменной. Для применения

рассмотренного алгоритма была использована информация с реального объекта – автоклавной установки для производства композитных элементов авиатехники.

Ключевые слова: предиктивная аналитика, машинное обучение, линейная регрессия, автоклавная установка

Для цитирования: Гусев П.Ю., Таволжанский А.В., Поцербнева И.В. Моделирование выбора характеристик временных рядов для методов машинного обучения на основе корреляционной матрицы // Труды МАИ. 2024. № 138. URL: <https://trudymai.ru/published.php?ID=182676>

Original article

MODELING THE SELECTION OF TIME SERIES CHARACTERISTICS FOR MACHINE LEARNING METHODS BASED ON THE CORRELATION MATRIX

Pavel Yu. Gusev¹✉, Alexander V. Tavolzhanskij², Irina V. Potsebneva³

Voronezh State Technical University,

Voronezh, Russia

¹pgusev@cchgeu.ru✉

²tavolzhanskij.a@yandex.ru

Abstract. This paper considers the problem of selecting the necessary set of characteristics for time series forecasting. The method of solving this problem on the basis of correlation matrix is proposed. A correlation matrix is constructed using the prepared data, after which a list is formed for each variable, ordered by decreasing modulus of the correlation degree. Then linear regression models are trained and the quality of predictions for different sets of

variables from the sorted list is compared. Next, the comparison is performed for different values of sampling time to determine the optimal value for each variable. To apply the considered algorithm, information from a real object - an autoclave plant for the production of composite aircraft components - was used. The use of autoclave plants for the production of aircraft parts represents a crucial stage in the manufacturing process of modern aircraft. The application of predictive analysis in the management of these facilities can facilitate the optimization of production processes and the enhancement of product quality. This paper proposes a method for identifying a set of characteristics based on a correlation matrix. The application of this method allows the most significant characteristics for predicting the condition of an autoclave plant for the production of aircraft parts to be identified, and those that do not affect the accuracy of the prediction to be excluded. In addition to the theoretical justification, the paper investigates the prediction accuracy depending on the number of variables involved in the prediction, as well as the sampling time, taking into account the different frequency of obtaining measurements.

Keywords: predictive analytics, machine learning, linear regression, autoclave plant

For citation: Gusev P.Yu., Tavolzhanskij A.V., Potsebneva I.V. Modeling the selection of time series characteristics for machine learning methods based on the correlation matrix.

Trudy MAI, 2024, no. 138. URL: <https://trudymai.ru/eng/published.php?ID=182676>

Введение

В последние годы прогнозирование временных рядов стало одной из наиболее актуальных задач в области анализа данных [1-4]. Однако, несмотря на значительные успехи в этой области, многие задачи остаются без должного внимания со стороны

научного сообщества. Одной из таких задач является определение минимального набора характеристик, необходимых для прогнозирования временных рядов.

Для прогнозирования временных рядов часто используются предварительно определенные целевые показатели и соответствующие данные [5]. Выбор показателей и данных основан на требованиях технологического процесса и фактическом количестве измерительных устройств. Однако, у этого подхода есть некоторые ограничения. Одним из них является эмпирическое определение необходимого набора данных, которое может ухудшить качество результатов при большом количестве переменных. Кроме того, существуют случаи, когда величину и поведение целевого показателя невозможно точно спрогнозировать по имеющимся данным. Часто, в таких случаях прибегают к использованию аппарата нейронных сетей [6-10].

Автоклавные установки для производства деталей авиатехники являются важным элементом в процессе создания современных самолетов. Они позволяют производить высококачественные композитные элементы, которые обладают высокой прочностью и легкостью. Использование предиктивного анализа в управлении этими установками может помочь оптимизировать процесс производства и повысить качество продукции.

В данной работе предлагается способ, позволяющий определить набор характеристик на основе корреляционной матрицы. Применение данного способа позволяет выявить наиболее значимые характеристики для прогнозирования состояния автоклавной установки [11-12] для производства частей авиатехники и исключить те, которые не влияют на точность прогноза. Помимо теоретического

обоснования, в работе проводится исследование точности прогнозирования в зависимости от количества переменных, участвующих в прогнозе, а также времени дискретизации, учитывая различную частоту получения измерений.

Постановка задачи и планирование эксперимента

Предиктивный анализ работы оборудования является важным этапом планирования, который позволяет прогнозировать возможные проблемы и оптимизировать работу предприятия [13-13]. С помощью предиктивного анализа можно определять потенциальные риски и уязвимости в работе оборудования, а также выявлять причины их возникновения. Такие сведения позволяют предотвратить возможные поломки и аварии, а также сократить затраты на ремонт и обслуживание [15-**Error! Reference source not found.**].

Кроме того, использование предиктивного анализа способно оптимизировать работу оборудования, улучшая его эффективность и снижая затраты на энергию.

Рассмотрим аспекты постановки исследуемой задачи.

Во-первых, требуется получить прогноз для целевого показателя на основе измерений оставшихся показателей, исключая собственные измерения.

Во-вторых, необходима возможность разбивать переменные на два набора: те, для которых прогнозирование возможно с приемлемой точностью, и те, для которых это невозможно.

В-третьих, для первого набора необходимо определить признаки, необходимые для прогнозирования, а также оптимальные условия для обучения модели.

В качестве входных данных используются показания измерительных приборов с автоклавной установки. Целью исследования является анализ величины средней квадратической ошибки (MSE) в следующих условиях [0]:

При различной глубине косвенной корреляции;

При различном количестве исходных переменных;

При различном количестве переменных, используемых для прогнозирования;

При различном времени дискретизации.

Параметры экспериментов

Таблица 1

Эксперимент	1	2	3	4
Количество исходных переменных, n	13	23		
Шаг дискретизации, s, δ	5			3, 5, 10
Количество значений для строки временного ряда, q	10			
Срок прогнозирования, w	500			
Количество используемых переменных, \hat{n}	12, 6	22, 12, 6	22, 12, 8, 6, 5, 3	8, 6, 5
Количество переменных для косвенной корреляции, γ	0, 1, 2		1	
Количество исходных переменных, участвующих в косвенной корреляции, ζ	1, 2, 12	1, 2, 22	1	

Методика проведения исследования

Пусть исходные данные представлены в следующем виде (табл. 2):

Исходные данные

Таблица 2

	F_1	F_2	...	F_n
DT_1	$V_{1,1}$	$V_{1,2}$...	$V_{1,n}$
DT_2	$V_{2,1}$	$V_{2,2}$...	$V_{2,n}$
...
DT_m	$V_{m,1}$	$V_{m,2}$...	$V_{m,n}$

где DT_m – момент времени, F_n – показатель, $V_{m,n}$ – значение показателя F_n в момент времени T_m .

Тогда корреляционную матрицу R по табл. 2 представим в виде (1):

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{pmatrix} \quad (1)$$

где r_{ij} – коэффициенты корреляции.

Далее сформируем новую матрицу \hat{R} , элементы которой представляют собой элементы матрицы R , взятые по модулю (2):

$$\hat{R}_{ij} = |R_{ij}|, \quad i, j \in [1, n], \quad i, j, n \in \mathbb{N} \quad (2)$$

Затем, для каждой переменной формируем список K , представляющий собой отсортированный столбец матрицы \hat{R} по убыванию, где величина корреляции заменяется названием характеристики.

Предлагается рассмотреть два подхода к выбору переменных.

1. Список переменных для прогнозирования формируется путем отбора k первых элементов из списка K . Однако, для повышения точности прогноза, предлагается использовать метод определения набора параметров с учетом косвенной корреляции [19-20].

2. Суть метода заключается в добавлении к уже отобранному списку k параметров определенного количества γ косвенно коррелирующих с ним переменных. Это может быть применено ко всем отобранными переменными, либо

только к одной или двум наиболее коррелирующим с целевым показателем признакам. Число таких переменных определим как ζ .

Предположим, что мы имеем дело с целевым показателем F_x , который обозначим как T . Исходя из этого, структура временного ряда для обучения будет содержать данные, собранные за z временных промежутков, где в каждом промежутке учитываются количество дискретных значений q для \hat{n} различных показателей, исключая параметр T . При этом целевой показатель смещается относительно текущего времени на w временных промежутков. При этом шаг дискретизации определяется, как:

$$DT_{a+1} = DT_a + \delta, a \in [k, zk] \quad (3)$$

Тогда преобразованная структура исходных из таблицы 2 примет вид, показанный в таблице 3.

Полученная структура данных

Таблица 3

	$1F_1$	$2F_1$...	KF_1	...	$1F_j$	$2F_j$...	KF_j	T
DT_k	$V_{1,1}$	$V_{2,1}$...	$V_{k,1}$...	$V_{1,j}$	$V_{2,j}$...	$V_{k,j}$	$V_{kw,1}$
DT_{2k}	$V_{k+1,1}$	$V_{k+2,1}$...	$V_{2k,1}$...	$V_{k+1,j}$	$V_{k+2,j}$...	$V_{2k,j}$	$V_{(k+1)w,1}$
...
DT_{zk}	$V_{(z-1)k+1,1}$	$V_{(z-1)k+2,1}$...	$V_{zk,1}$...	$V_{(z-1)k+1,j}$	$V_{(z-1)k+2,j}$...	$V_{zk,j}$	$V_{zkw,1}$

Затем, сформированная структура данных будет использована для обучения линейной регрессионной модели. Линейная регрессия является базовой моделью машинного обучения, которая служит основой для более сложных моделей. Она позволяет оценить влияние отдельных факторов на зависимую переменную, обеспечивая интерпретируемость результатов.

В качестве целевой функции обучения модели используется среднеквадратичная ошибка (MSE).

Для оценки точности прогнозов, полученных при различных условиях, будем использовать MSE, поскольку этот метод более чувствителен к значительным отклонениям, что особенно важно при анализе внештатных ситуаций.

Проведение экспериментов и анализ полученных результатов

В качестве исходных данных рассматриваются показания измерительных приборов автоклавной установки вместе со станцией водоподготовки. Данные представляют собой записи девяти технологических цикла изготовления композитных делателей авиатехники.

Для обучения модели используются данные первого цикла, после чего модель тестируется на всех остальных показаниях, с последующим усреднением среднеквадратичная ошибки для всех девяти циклов.

В рамках первого эксперимента мы проведем исследование среднеквадратичной ошибки при различных значениях глубины косвенной корреляции. Это позволит нам оценить влияние данного параметра на точность моделирования (табл. 4).

Результаты эксперимента №1 (MSE)

Таблица 4

Эксперимент	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
	12UID	6UID	Для одной переменной ($\zeta = 1$)		Для двух переменных ($\zeta = 2$)		Для всех переменных ($\zeta = 12$)	
γ	0	0	1	2	1	2	1	2
UID1	17,7125	11,8997	11,9143	11,9711	11,9711	11,3495	11,9711	11,3495
UID2	2,043E-05	3,479E-05	3,431E-05	3,488E-05	3,488E-05	3,494E-05	3,488E-05	3,494E-05
UID3	0,1962	0,2132	0,2265	0,2267	0,2267	0,2255	0,2153	0,2255
UID4	4,465E-04	4,461E-04	4,709E-04	4,564E-04	4,629E-04	5,135E-04	4,567E-04	5,135E-04

UID5	2,727E-04	3,637E-04	3,639E-04	4,394E-04	4,394E-04	3,714E-04	4,394E-04	3,714E-04
UID6	164,5231	164,4332	164,6342	159,8769	165,2903	158,9253	158,4562	158,9253
UID7	36,5125	36,4552	36,4552	36,5194	35,7301	44,9435	35,9917	44,9435
UID8	1,193E-02	9,662E-03	9,662E-03	1,129E-02	1,108E-02	1,185E-02	1,323E-02	1,185E-02
UID9	1,034E-03	1,031E-03	1,133E-03	1,063E-03	1,131E-03	9,568E-04	1,131E-03	9,568E-04
UID10	2,052E-01	2,131E-01	2,065E-01	2,026E-01	2,026E-01	2,529E-01	2,005E-01	2,529E-01
UID11	12,3346	12,3235	12,3654	12,4390	12,4390	7,2828	12,2481	7,2828
UID12	1,8232	1,5104	1,4188	1,6518	1,4365	1,5763	1,6925	1,5763
UID13	464,9015	456,2357	475,1887	478,1888	478,3472	498,5845	478,3472	498,5845

В таблицах 4-7: ячейка с зеленым выделением – минимальная ошибка для признака, красная – максимальная ошибка.

Результаты первого эксперимента показывают, что для большинства UID ошибка по сравнению с эталонным экспериментом 1.1 снижается. Также видно, что при увеличении количества добавляемых переменных γ , а также с ростом количества используемых ζ ошибка растёт. Одинаковые результаты для $\zeta = [2, 12]$ связаны с малым количеством переменных, соответственно списки отобранных признаков одинаковы для некоторых UID.

Далее проведём аналогичный эксперимент, но с увеличением количества исходных переменных с 13 до 23 (табл. 5). Проведём сравнение результатов с использованием всех переменных 22UID, половины переменных 12UID, а также с шестью UID для сравнения как с первым экспериментом, так и с остальными случаями в рамках второго.

Эксперимент	2.1	2.2	2.3	2.4		2.5	2.6		2.7	2.8	2.9
	22UID	12UID	6UID	Для одной переменной ($\zeta = 1$)			Для двух переменных ($\zeta = 2$)			Для всех переменных ($\zeta = 22$)	
γ	0	0	0	1	2		1	2		1	2
UID1	21,9251	19,4801	29,3438	29,3438	29,3438		29,3438	21,8272		30,1484	44,2793
UID2	3,27E-05	3,61E-05	2,74E-05	2,73E-05	2,73E-05		2,74E-05	2,86E-05		2,78E-05	1,93E-04
UID3	0,1729	0,1722	0,1689	0,1689	0,1702		0,1689	0,1715		0,1695	0,1711
UID4	4,05E-04	3,65E-04	4,00E-04	4,00E-04	4,16E-04		4,00E-04	4,00E-04		4,00E-04	3,99E-04
UID5	2,09E-04	3,81E-04	2,22E-04	2,22E-04	2,22E-04		2,22E-04	1,58E-04		1,65E-04	1,58E-04
UID6	140,3576	146,4357	137,7209	142,7147	144,5696		142,6531	148,8502		145,8619	148,8502
UID7	33,7028	35,8976	36,6931	36,1448	38,2179		40,1832	37,5644		39,4177	40,9000
UID8	0,0087	0,0090	0,0084	0,0084	0,0085		0,0084	0,0092		0,0091	0,0076
UID9	1,12E-03	1,37E-03	1,33E-03	1,33E-03	9,47E-04		9,50E-04	8,33E-04		9,97E-04	7,26E-04
UID10	0,5232	0,3554	0,4905	0,4905	0,4416		0,4416	0,4510		0,4143	0,4510
UID11	7,4523	10,1934	6,0139	6,0139	5,0011		6,4733	2,0407		6,4941	3,2699
UID12	13,4273	9,2192	0,8823	10,1041	8,4208		8,4208	7,2524		12,3142	7,2524
UID13	434,0808	446,4379	409,8250	409,8250	435,3566		439,5382	432,1767		445,6011	432,1767
UID14	4,4989	5,7393	5,4793	5,4793	5,1002		5,4793	5,2943		5,4793	4,1589
UID15	2,0182	2,2962	1,2002	1,2002	1,2002		1,2002	1,0410		1,0410	0,8727
UID16	0,0031	0,0034	0,0025	0,0025	0,0025		0,0025	0,0022		0,0022	0,0014
UID17	7,27E-04	7,66E-04	7,15E-04	7,72E-04	7,67E-04		7,67E-04	7,67E-04		8,32E-04	7,86E-04
UID18	0,0263	0,0253	0,0310	0,0310	0,0309		0,0309	0,0534		0,0534	0,0286
UID19	6,94E-06	6,74E-06	3,46E-06	3,46E-06	3,53E-06		3,53E-06	3,97E-06		3,53E-06	3,97E-06
UID20	69,9759	58,1114	34,5650	34,5650	34,5650		34,5650	49,5226		34,5650	12,4491
UID21	8,0063	5,5163	13,6467	13,6467	13,5971		13,5971	13,5971		14,6681	13,5971
UID22	0,2487	0,2447	0,1919	0,2775	0,2831		0,2831	0,2825		0,2981	0,2465
UID23	329,1013	310,8012	860,3734	860,3734	845,3771		836,6039	446,1616		836,6039	435,1362

Второй эксперимент иллюстрирует следующее: использование всех переменных, а также простое исключение половины признаков показывают результаты хуже, чем предлагаемый метод (эксперименты 2.1 и 2.2). То есть хорошие результаты экспериментов 1.1 и 1.2 связаны с малым количеством исходных переменных.

Для дальнейшего исследования будут использованы параметры эксперимента 2.4, а именно $\zeta = 1$, где будет исследована ошибка для разного количества используемых признаков \hat{n} (табл. 6).

Эксперимент	3.1	3.2	3.3	3.4	3.5	3.6
\hat{n}	22UID	12UID	8UID	6UID	5UID	3UID
UID1	21,9251	19,4801	18,6285	29,3438	29,3438	38,5931
UID2	3,27E-05	3,61E-05	3,64E-05	2,74E-05	2,74E-05	3,07E-05
UID3	0,1729	0,1722	0,1806	0,1689	0,1702	0,1707
UID4	4,05E-04	3,65E-04	3,89E-04	4,00E-04	4,16E-04	4,89E-04
UID5	2,09E-04	3,81E-04	3,69E-04	2,22E-04	2,22E-04	1,58E-04
UID6	140,3576	146,4357	142,5102	142,7147	142,6531	145,2973
UID7	33,7028	35,8976	36,0262	36,1448	39,6931	39,1850
UID8	0,0087	0,0090	0,0088	0,0084	0,0085	0,0086
UID9	1,12E-03	1,37E-03	1,35E-03	1,33E-03	9,95E-04	9,71E-04
UID10	0,5232	0,3554	0,4962	0,4905	0,4416	0,4279
UID11	7,4523	10,1934	6,6301	6,0139	6,0139	6,7104
UID12	13,4273	9,2192	8,9920	10,1041	8,4208	5,5121
UID13	434,0808	446,4379	408,8020	409,8250	439,5382	440,5795
UID14	4,4989	5,7393	5,6559	5,4793	5,4793	6,0817
UID15	2,0182	2,2962	1,2118	1,2002	1,2002	2,1915
UID16	0,0031	0,0034	0,0023	0,0025	0,0025	0,0059
UID17	7,27E-04	7,66E-04	7,59E-04	7,72E-04	7,67E-04	8,37E-04
UID18	0,0263	0,0253	0,0265	0,0310	0,0309	0,0562
UID19	6,94E-06	6,74E-06	3,38E-06	3,46E-06	3,53E-06	2,70E-06
UID20	69,9759	58,1114	32,4097	34,5650	34,5650	47,7579
UID21	8,0063	5,5163	12,1952	13,6467	13,5971	15,9503
UID22	0,2487	0,2447	0,2870	0,2775	0,2831	0,5510
UID23	329,1013	310,8012	325,7551	860,3734	874,1462	128,6250

В рамках эксперимента 3.2 наблюдаются улучшение результатов по сравнению с экспериментом 2. Чрезмерное сокращение объема исходных данных негативно влияет на точность прогнозирования.

Результаты для 3.3-3.6 превосходят показатели других исследований. Параметры в данном диапазоне будут подвергнуты дальнейшему изучению при различных временных интервалах дискретизации, отличающихся от текущего времени.

Поскольку признаки характеризуются различной частотой дискретизации от ~1,5 до ~20 минут, планируется проведение эксперимента со следующими интервалами дискретизации: 3, 5 и 10 секунд (табл. 7).

Эксперимент	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9
δ, c	3			5			10		
\hat{n}	8UID	6UID	5UID	8UID	6UID	5UID	8UID	6UID	5UID
UID1	21,7489	32,0017	32,0017	18,6285	29,3438	29,3438	13,8580	30,5874	30,5874
UID2	3,21E-05	2,67E-05	2,67E-05	3,64E-05	2,74E-05	2,74E-05	8,91E-05	4,19E-05	4,19E-05
UID3	0,2366	0,2395	0,2279	0,1806	0,1689	0,1702	0,4516	0,4394	0,4408
UID4	3,86E-04	3,84E-04	3,84E-04	3,89E-04	4,00E-04	4,16E-04	3,96E-04	3,85E-04	3,88E-04
UID5	2,51E-04	2,52E-04	1,72E-04	3,69E-04	2,22E-04	2,22E-04	6,72E-04	4,30E-04	2,81E-04
UID6	144,8062	144,2691	144,2294	142,5102	142,7147	142,6531	150,6975	149,5361	149,6116
UID7	35,8574	35,4301	35,3815	36,0262	36,1448	39,6931	38,9600	39,5227	38,7386
UID8	0,0091	0,0099	0,0082	0,0088	0,0084	0,0085	0,0133	0,0133	0,0109
UID9	1,12E-03	1,05E-03	1,06E-03	1,35E-03	1,33E-03	9,95E-04	9,92E-04	1,03E-03	1,89E-03
UID10	0,4890	0,4822	0,4063	0,4962	0,4905	0,4416	0,5831	0,5492	0,5401
UID11	5,3425	5,3884	5,3884	6,6301	6,0139	6,0139	9,6795	9,2327	9,2327
UID12	0,7874	0,8137	1,5397	8,9920	10,1041	8,4208	17,9856	25,5719	20,9965
UID13	381,6416	385,4735	419,4787	408,8020	409,8250	439,5382	568,6195	573,6060	522,7668
UID14	5,6690	5,8360	5,8360	5,6559	5,4793	5,4793	7,0997	5,6031	5,6031
UID15	1,2475	1,2171	1,2171	1,2118	1,2002	1,2002	2,2430	2,2641	2,2641
UID16	0,0015	0,0016	0,0016	0,0023	0,0025	0,0025	0,0052	0,0054	0,0054
UID17	9,51E-04	8,75E-04	8,61E-04	7,59E-04	7,72E-04	7,67E-04	1,40E-03	1,50E-03	8,83E-04
UID18	0,0288	0,0352	0,0351	0,0265	0,0310	0,0309	0,0256	0,0230	0,0231
UID19	3,80E-06	3,79E-06	3,60E-06	3,38E-06	3,46E-06	3,53E-06	3,19E-06	3,07E-06	3,02E-06
UID20	19,6249	21,3858	21,3858	32,4097	34,5650	34,5650	3,9206	2,6399	2,6399
UID21	11,5883	12,6385	14,5578	12,1952	13,6467	13,5971	16,7125	17,7916	11,8957
UID22	0,5905	0,5663	0,5740	0,2870	0,2775	0,2831	1,2787	1,2741	1,2793
UID23	289,9719	554,8795	555,5768	325,7551	860,3734	874,1462	703,8843	1023,9710	1030,6059

В ходе проведения четвертого эксперимента было установлено, что для большого количества переменных уменьшение временного интервала между измерениями приводит к улучшению качества данных. Кроме того, были зафиксированы случаи, когда увеличение шага измерений до 10 секунд сопровождалось повышением точности прогнозирования. Это свидетельствует о том, что изменение шага дискретизации может существенно влиять на конечный результат.

Для сравнительного анализа эффективности метода построим сводную таблицу с наилучшими результатами без использования (эксперименты 2.1-2.3) и с использованием предложенного метода (эксперименты 2.4-4.9) (табл. 8).

Сводная таблица результатов (MSE)

Таблица 8

	Без использования метода	С использованием метода	Разница
UID1	19,4801	13,858	-28,86%
UID2	2,74E-05	2,67E-05	-2,55%
UID3	0,1689	0,1689	0%
UID4	3,65E-04	3,65E-04	0%
UID5	2,09E-04	1,58E-04	-24,40%
UID6	137,7209	142,5102	3,48%
UID7	33,7028	35,3815	4,98%
UID8	0,0084	0,0076	-9,52%
UID9	0,00112	0,000726	-35,18%
UID10	0,3554	0,3554	0%
UID11	6,0139	2,0407	-66,07%
UID12	0,8823	0,7874	-10,76%
UID13	409,825	381,6416	-6,88%
UID14	4,4989	4,1589	-7,56%
UID15	1,2002	0,8727	-27,29%
UID16	0,0025	0,0014	-44,00%
UID17	0,000715	0,000759	6,15%
UID18	0,0253	0,023	-9,09%
UID19	3,46E-06	0,0000027	-21,97%
UID20	34,565	2,6399	-92,36%
UID21	5,5163	5,5163	0%
UID22	0,1919	0,2447	27,51%
UID23	310,8012	128,625	-58,62%

Заключение

Анализ данных, представленных в сводной таблице, демонстрирует, что предложенная методика способствует улучшению качества прогнозов для большинства переменных. Однако для некоторых признаков наблюдается как сохранение прежних результатов, так и снижение точности прогнозирования. Это может быть связано с особенностями структуры данных, спецификой исследуемых процессов или другими факторами, требующими дополнительного изучения. Важно отметить, что несмотря на эти исключения, в целом предложенная методика

показывает свою эффективность и может быть рекомендована для использования в аналогичных исследованиях.

Список источников

1. Ковалев В.З., Швецов С.Ю., Архипова О.В. Анализ методов прогнозирования потребления электрической энергии и мощности // Инженерный вестник Дона. 2023. № 4 (100). С. 127-141.

2. Неизвестный О.Г. Программа имитационного моделирования системы предиктивного управления процессом дегидрирования этилбензола // Вестник Воронежского государственного технического университета. 2023. Т. 19. № 6. С. 32-40. DOI: [10.36622/VSTU.2023.19.6.005](https://doi.org/10.36622/VSTU.2023.19.6.005)

3. Shehadeh A., Alshboul O., Al Mamlook R. E., Hamedat O. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression // Automation in Construction, 2021, vol. 129, no. 2, pp. 103827. DOI: [10.1016/j.autcon.2021.103827](https://doi.org/10.1016/j.autcon.2021.103827)

4. Семаков С.Л., Семаков И.С. Простейшая прогнозная модель временного ряда и ее реакция на линейное и параболическое входные воздействия // Труды МАИ. 2018. № 100. URL: <https://trudymai.ru/published.php?ID=93446>

5. Седых И.А., Стругов И.В. Определение оптимального набора переменных при помощи методов машинного обучения при исследовании рейтинговых систем // Вестник Воронежского государственного технического университета. 2023. Т. 19. № 2. С. 72-78.

6. Черников А.А. Алгоритм обнаружения и классификации объектов на неоднородном фоне для оптико-электронных систем // Труды МАИ. 2023. № 129. URL: <https://trudymai.ru/published.php?ID=173039>. DOI: [10.34759/trd-2023-129-26](https://doi.org/10.34759/trd-2023-129-26)
7. Чигринец Е.Г., Верченко А.В. CAD/CAM/CAE системы, OMW-технологии и нейросетевые алгоритмы анализа данных на предприятиях авиастроительной отрасли // Труды МАИ. 2019. № 104. URL: <https://trudymai.ru/published.php?ID=102420>
8. Manoharan A., Begam K.M., Aparow V.R., Sooriamoorthy D. Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review // Journal of Energy Storage, 2022, no. 55, pp. 105384. DOI: [10.1016/j.est.2022.105384](https://doi.org/10.1016/j.est.2022.105384)
9. James G., Witten D., Hastie T., Tibshirani R., Taylor J. An introduction to statistical learning: With applications in python, Springer Nature, 2023, 607 p.
10. Румянцев Н.В., Соловьёв С.В., Павлов Д.В. Интеллектуализированная система контроля, диагностики и прогнозирования состояния бортовых систем космических аппаратов // Труды МАИ. 2024. № 136. URL: <https://trudymai.ru/published.php?ID=180688>
11. Чижов М.И., Скрипченко Ю.С., Гусев П.Ю. Имитационное моделирование производства деталей из полимерных композиционных материалов // Компьютерные исследования и моделирование. 2014. Т. 6. № 2. С. 245-252.
12. Гусев П.Ю., Скрипченко Ю.С., Лысов Д.В. Анализ и оптимизация производства деталей самолетных агрегатов с применением имитационного моделирования // Известия Самарского научного центра Российской академии наук. 2016. Т. 18. № 4-3. С. 432-438.

13. Гусев П.Ю. Автоматизация планирования производственных процессов авиастроительного предприятия с использованием цифрового двойника // Труды МАИ. 2018. № 103. URL: <https://trudymai.ru/published.php?ID=101190>
14. Ramesh T.R., Lilhore U.K., Poongodi M., Simaiya S., Kaur A., Hamdi M. Predictive analysis of heart diseases with machine learning approaches // Malaysian Journal of Computer Science, 2022, no. 1, pp. 132-148. DOI: [10.22452/mjcs.sp2022no1.10](https://doi.org/10.22452/mjcs.sp2022no1.10)
15. Таволжанский А.В., Гусев П.Ю. Инструменты предиктивной аналитики // 12-я Международная молодежная научная конференция «Поколение будущего: взгляд молодых ученых-2023» (Курск, 09–10 ноября 2023): сборник статей. – Курск: Университетская книга, 2023. Т. 3. С. 96-99.
16. Гусев П.Ю., Таволжанский А.В. Алгоритмизация обработки и подготовки данных для построения моделей предиктивной аналитики // Вестник Воронежского государственного технического университета. 2024. Т. 19. № 2. С. 14-19.
17. Gusev P.Yu., Tavolzhanskij A.V. Techniques and algorithms for predictive analysis // Международный форум профессионального образования «Антропоцентрические науки в образовании: вызовы, трансформации, ресурсы» (Воронеж, 09–10 апреля 2024): сборник статей. – Воронеж: Научная книга, 2024. С. 396-400.
18. Qi J., Du J., Siniscalchi S.M., Ma X., Lee C.H. On mean absolute error for deep neural network based vector-to-vector regression // IEEE Signal Processing Letters, 2020, no. 27, pp. 1485-1489. DOI: [10.1109/LSP.2020.3016837](https://doi.org/10.1109/LSP.2020.3016837)
19. Ризванова Э.Р. Корреляционная матрица в анализе статистической информации структуры внешней торговли // Известия Санкт-Петербургского государственного экономического университета. 2017. № 6 (108). С. 159-163.

20. Егошин В.Л., Иванов С.В., Саввина Н.В., Ермолаев А.Р., Мамырбекова С.А., Жамалиева Л.М., Гржибовский А.М. Корреляционный и простой линейный регрессионный анализ с использованием программной среды r // Экология человека. 2018. № 12. С. 55-64.

References

1. Kovalev V.Z., Shvetsov S.Yu., Arkhipova O.V. *Inzhenernyi vestnik Dona*, 2023, no. 4 (100), pp. 127-141.
2. Neizvestnyi O.G. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta*, 2023, vol. 19, no. 6, pp. 32-40. DOI: [10.36622/VSTU.2023.19.6.005](https://doi.org/10.36622/VSTU.2023.19.6.005)
3. Shehadeh A., Alshboul O., Al Mamlook R. E., Hamedat O. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression, *Automation in Construction*, 2021, vol. 129, no. 2, pp. 103827. DOI: [10.1016/j.autcon.2021.103827](https://doi.org/10.1016/j.autcon.2021.103827)
4. Semakov S.L., Semakov I.S. *Trudy MAI*, 2018, no. 100. URL: <https://trudymai.ru/eng/published.php?ID=93446>
5. Sedykh I.A., Strugov I.V. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta*, 2023, vol. 19, no. 2, pp. 72-78.
6. Chernikov A.A. *Trudy MAI*, 2023, no. 129. URL: <https://trudymai.ru/eng/published.php?ID=173039>. DOI: [10.34759/trd-2023-129-26](https://doi.org/10.34759/trd-2023-129-26)
7. Chigrinets E.G., Verchenko A.V. *Trudy MAI*, 2019, no. 104. URL: <https://trudymai.ru/eng/published.php?ID=102420>

8. Manoharan A., Begam K.M., Aparow V.R., Sooriamoorthy D. Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review, *Journal of Energy Storage*, 2022, no. 55, pp. 105384. DOI: [10.1016/j.est.2022.105384](https://doi.org/10.1016/j.est.2022.105384)
9. James G., Witten D., Hastie T., Tibshirani R., Taylor J. *An introduction to statistical learning: With applications in python*, Springer Nature, 2023, 607 p.
10. Rumyantsev N.V., Solov'ev S.V., Pavlov D.V. *Trudy MAI*, 2024, no. 136. URL: <https://trudymai.ru/eng/published.php?ID=180688>
11. Chizhov M.I., Skripchenko Yu.S., Gusev P.Yu. *Komp'yuternye issledovaniya i modelirovanie*, 2014, vol. 6, no. 2, pp. 245-252.
12. Gusev P.Yu., Skripchenko Yu.S., Lysov D.V. *Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk*, 2016, vol. 18, no. 4-3, pp. 432-438.
13. Gusev P.Yu. *Trudy MAI*, 2018, no. 103. URL: <https://trudymai.ru/eng/published.php?ID=101190>
14. Ramesh T.R., Lilhore U.K., Poongodi M., Simaiya S., Kaur A., Hamdi M. Predictive analysis of heart diseases with machine learning approaches, *Malaysian Journal of Computer Science*, 2022, no. 1, pp. 132-148. DOI: [10.22452/mjcs.sp2022no1.10](https://doi.org/10.22452/mjcs.sp2022no1.10)
15. Tavolzhanskii A.V., Gusev P.Yu. *12-ya Mezhdunarodnaya molodezhnaya nauchnaya konferentsiya «Pokolenie budushchego: vzglyad molodykh uchenykh-2023»*: sbornik statei. Kursk, Universitetskaya kniga, 2023, vol. 3, pp. 96-99.
16. Gusev P.Yu., Tavolzhanskii A.V. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta*, 2024, vol. 19, no. 2, pp. 14-19.

17. Gusev P.Yu., Tavolzhanskij A.V. Techniques and algorithms for predictive analysis, *Mezhdunarodnyi forum professional'nogo obrazovaniya «Antropotsentricheskie nauki v obrazovanii: vyzovy, transformatsii, resursy»* sbornik statei. Voronezh, Nauchnaya kniga, 2024, pp. 396-400.
18. Qi J., Du J., Siniscalchi S.M., Ma X., Lee C.H. On mean absolute error for deep neural network based vector-to-vector regression, *IEEE Signal Processing Letters*, 2020, no. 27, pp. 1485-1489. DOI: [10.1109/LSP.2020.3016837](https://doi.org/10.1109/LSP.2020.3016837)
19. Rizvanova E.R. *Izvestiya Sankt-Peterburgskogo gosudarstvennogo ekonomicheskogo universiteta*, 2017, no. 6 (108), pp. 159-163.
20. Egoshin V.L., Ivanov S.V., Savvina N.V., Ermolaev A.R., Mamyrbekova S.A., Zhamalieva L.M., Grjibovskij A.M. *Ekologiya cheloveka*, 2018, no. 12, pp. 55-64.

Статья поступила в редакцию 13.08.2024

Одобрена после рецензирования 01.09.2024

Принята к публикации 25.10.2024

The article was submitted on 13.08.2024; approved after reviewing on 01.09.2024; accepted for publication on 25.10.2024