ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 4 (2), 2024

https://doi.org/10.23939/ictee2024.02.010

R^2 METRIC DYNAMICS FOR K-NEAREST NEIGHBORS REGRESSION MODEL TRAINED ON SERIES OF DIFFERENT SIZES

Y. Babich [ORCID: 0000-0002-7888-7591], **L. Hlazunova,**
T. Kalinina [ORCID: 0000-0002-3184-3604], **Y. Petrovych** [ORCID: 0009-0008-8939-2333]

State University of Intellectual Technologies and Telecommunications, 1, Kuznechna str., Odesa, 65023, Ukraine

Corresponding author: Y. Babich (e-mail: y.o_babich@suitt.edu.ua).

(Received 20 March 2024)

An R2 score or a coefficient of determination is used often as a metric to evaluate regression models. It can be applied solely but usually it is combined with other metrics in order to increase accuracy of a model evaluation. The goal of the work is to research the dynamics of the R2 score of a K-Nearest Neighbors regression model trained on series of different sizes in order to propose a new approach to increase the robustness and accuracy of the model evaluation when the R2 score metric is used solely. Typically, a value of the R2 score metric above 0.8 is considered to be sufficient while an evaluated model is considered to be accurate enough. However, such a way of R2 score interpretation to may lead to model's accuracy misevaluation, which is shown in the proposed paper. The results obtained clearly display that R2 score can vary significantly in some cases depending on the samples selected to test part of a series used for model evaluation. The mentioned variation can contribute to model's accuracy overestimation, which, in turn can lead to incorrect results of model application. The known methods to make model estimation more accurate involve use of other metrics. Instead, this paper focuses on increase of model's accuracy estimation without the necessity of using other metrics. The R2 score dynamics is examined using 25000 cycles of the K-Nearest Neighbors regression model training and evaluation. Selection of samples to a training or test part of a series has been done randomly. For all the experiments quantity of neighbors is fixed and equals to the default value of n_neighbors=5 of the KNeighborsRegressor method provided by the Sklearn library. The paper both states and proves a hypothesis that the R2 score variation is expected to increase with series size reduction and the variation is supposed to be observed for models trained on the same series because of training/test samples selection randomness. The experiments carried out allowed to propose an alternative approach that did not require any supplementary metrics. The proposed approach considers application of the R2 score along with its variation that must not exceed 0.2 for the K-Nearest Neighbors regression model.

Keywords: *series size, R^2 score, coefficient of determination, regression model.*
UDC: 004.8.

1. Introduction

Business-related data analysis usually includes tasks to predict (extrapolate) or interpolate some necessary values (for example, sales dynamics, capitalization, load on servers etc.). These tasks can be solved with the help of machine learning algorithms through training a regression model.

Significant step of data analysis is a step of model evaluation. Evaluation is typically carried out using metrics suitable for regression models.

Coefficient of determination or an R^2 score is a metric commonly used to evaluate the performance of regression models. This metric can be used alone or in a combination with Mean Square Error (MSE), Mean Absolute Error (MAE) or Mean Error (ME) to compare several models or to estimate accuracy of a single regression model.

According to [1], the R^2 score measures the proportion of variance in the dependent variable which is explained by the independent variable and can be calculated as:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \overline{y_i})^2}, \tag{1}$$

where y_i denotes an actual value of the dependent variable; $\overline{y}$ refers to a mean value of the dependent variable and $\hat{y}$ denotes a predicted value of the dependent variable.

It is seen form (1) that random selection of samples to training and test series for the same machine learning algorithm may result in variation of R^2 scores of regression models trained on the same dataset, which may lead to incorrect model evaluation.

The object of study is the process of applying the R^2 score metric to evaluate the K-Nearest Neighbors regression model trained on series of different sizes. Precisely, it is the R^2 score variation caused by series size changes and the randomness of selecting samples to training and test parts of series.

The purpose of the work is to propose a method to increase robustness and accuracy of an R^2 metric value interpretation for a K-Nearest Neighbors (KNN) regression model trained on series of different sizes without using MSE, MAE, ME or other suitable metrics.

2. Task statement

Let us formulate hypothesis about the R^2 score variation caused by series size reduction and the randomness of selecting samples to training and test parts of series. The R^2 score variation is expected to increase with series size reduction and the variation is supposed to be observed for models trained on the same series because of training/test samples selection randomness.

The proof of the hypothesis and understanding the R^2 score dynamics allows to propose a new method for R^2 score metric application without the necessity to use other metrics to increase accuracy and robustness of a KNN regression model evaluation. The new method can be based on limiting the R^2 score variation.

3. Related works

According to [1] and [2], the R^2 score is supposed to be between 0 and 1, the closer to 1, the better the regression fit. Mathematically it is expressed as:

$$R^2(y,\hat{y}) \to 1 \tag{2}$$

Despite being a well-described and a long-present metric, the R^2 score is a subject of actual researches. For example, the work [3] introduces a family of the R^2 score implementations for generalized linear mixed models considering different probability distributions. The work [4] proposes the R^2 score implementation for generalized linear models by means of using the variance function to define the total variation of the dependent variable, as well as the remaining variation of the dependent variable after modeling the predictive effects of the independent variables.

The common trend from [3] and [4] is the intention to widen the application of the R^2 score metric on its own. This approach is used in real scientific projects including [5, 6, 7].

However, the robustness and accuracy of a model evaluation can also be increased by combining the R^2 score with other metrics. The later approach can be seen in some scientific works. For example, it is used in the works [8, 9, 10].

To sum up, the scientific topics aimed at widening the area of the R^2 score metric application are relevant and are being developed now. Scientific projects involving the R^2 score metric solely for results estimations are common for different research areas. However, it is incorrect to state that the use of the R^2 score is completely determined and revealed for all possible cases of application. This makes relevant a search for new methods of the R^2 score sole application, which is the case of the present work.

4. Experiment setup

A series of samples must be generated in order to research the behavior of the R^2 score metric with the intention to prove or decline the hypothesis formulated above. Let us use the formula (3) to generate samples.

$$y = \begin{cases} \frac{1}{2}x,\ x \in [1; 16]; \\ 10\frac{7}{8}x - 166,\ x \in [16; 24]; \\ \frac{5}{28}x + 90\frac{5}{7},\ x \in [24; 52]; \\ -4\frac{1}{11}x + 312\frac{8}{11},\ x \in [52; 74]; \\ -\frac{5}{13}x + 38\frac{6}{13},\ x \in [74; 100]. \end{cases} \quad (3)$$

Let us define the range of an independent variable (x) that is used to generate a series for the KNN regression model training. In this particular experiment setup, we use the range of $x \in [1; 100]$ to generate the samples shown in the Fig.1.

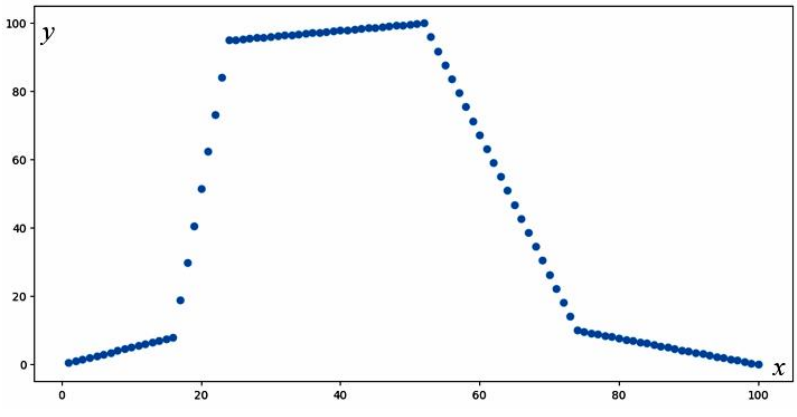

Fig. 1. Samples series for the hypothesis testing

The samples shown in the Fig.1 are fit with the K-Nearest Neighbors regression model. For each cycle of training/testing the R^2 score is computed and then series is reduced by 5 samples. Thus, series initially contains 100 samples, then 95, then is reduced to 90 etc., until the KNN model is underfit.

For each sample size (N= 100, 95, 90, …) KNN is trained 1000 times in order to estimate variation of the R^2 score. Precisely, there are 1000 values of the determination coefficient obtained per each sample size, which allows to estimate variation (ΔR^2), maximum R^2 score ($R^2_{\max}$), minimum R^2 score ($R^2_{\min}$), and a median value of R^2 metric (R^2_{median}). The variation of the R^2 score is calculated as follows:

$$\Delta R^2 = R^2_{\max} - R^2_{\min} \quad (4)$$

In the experiments a model is considered to be underfit if at least one value of the R^2 score is zero or negative.

The experiments carried out considered two proportions of training/test samples distributions recommended by [1] and [11]. These proportions were 70/30 and 80/20 (i.e. 70% or 80% of samples for training and 30% or 20% for testing).

Instrumentwise, the experiments were carried out using the pandas library [12] for data structures and data preparation, the NumPy library [13] to reshape data before training the model, the KNN regression model was implemented by the KNeighborsRegressor [14] module from the Scikit-learn library and the R^2 score metric was calculated by means of the sklearn.metrics.r2_score [2] module from the Scikit-learn library. Graphs were built using the matplotlib.pyplot [15] library.

5. Experiments results

In the case of the 70/30 training/test samples distribution the experiment was stopped when the quantity of samples (*N*) reached 40. At this point the minimum value of the R^2 score was 0.016 and the variation of R^2 reached 0.9794. The experiment results for the case of the 70/30 training/test samples distribution are given in the Table 1.

Table 1

R2 score metric estimation for the case of the 70/30 training/test samples distribution.

70/30				
N	$R^2_{\max}$	$R^2_{\min}$	R^2_{median}	ΔR^2
100	0.9996	0.8981	0.9934	0.1015
95	0.9993	0.9186	0.9926	0.0807
90	0.999	0.8947	0.9922	0.1043
85	0.9988	0.836	0.9909	0.1628
80	0.999	0.8652	0.9905	0.1338
75	0.9984	0.8668	0.9881	0.1316
70	0.9986	0.8277	0.9861	0.1709
65	0.998	0.7974	0.9816	0.2006
60	0.9985	0.7373	0.9759	0.2612
55	0.9966	0.58	0.9648	0.4166
50	0.9961	0.6764	0.9597	0.3197
45	0.9955	0.388	0.956	0.6075
40	0.9954	0.016	0.944	0.9794

Table 2

R2 score metric estimation for the case of the 80/20 training/test samples distribution.

80/20				
N	$R^2_{\max}$	$R^2_{\min}$	R^2_{median}	ΔR^2
100	0.9998	0.9123	0.9955	0.0875
95	0.9997	0.8937	0.9951	0.106
90	0.9995	0.8706	0.9947	0.1289
85	0.9997	0.8737	0.9938	0.126
80	0.9995	0.9046	0.9938	0.0949
75	0.9995	0.8697	0.9916	0.1298
70	0.9996	0.7754	0.9902	0.2242
65	0.9993	0.8446	0.9886	0.1547
60	0.9994	0.0975	0.9852	0.9019
55	0.9983	0.4397	0.978	0.5586
50	0.999	0.0268	0.9743	0.9722
45	0.999	0	0.9662	0.999

In the case of the 80/20 training/test samples distribution the experiment was stopped when the quantity of samples (*N*) reached 45. At this point the minimum value of the R^2 score was 0 and the variation of R^2 reached 0.999. The experiment results for the case of the 80/20 training/test samples distribution are given in the Table 2.

Despite showing slightly different results, tables 1 and 2 demonstrate the same trend of the ΔR^2 score. Table 1 took 13000 training/testing cycles (1000 per each value of *N*) of the KNN regression model, while the Table 2 is based on 12000 of the cycles.

6. Results discussion

Fit accuracy of the KNN regression model decreases with the reduction of a series size, which is shown in the figures 2 and 3.

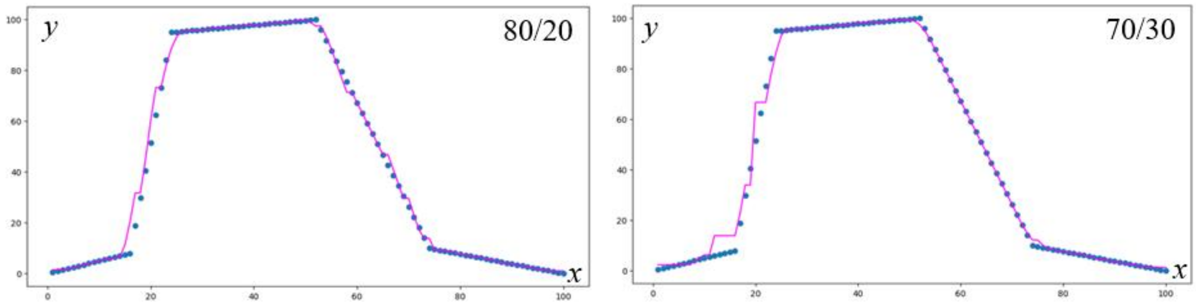

Fig. 2. KNN regression model (magenta line) fit to N=100 samples of data.

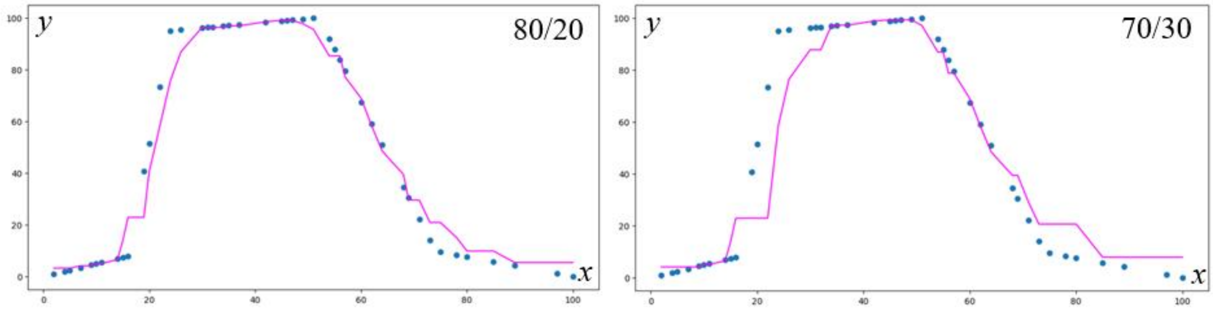

Fig. 3. KNN regression model (magenta line) fit to N=45 samples of data.

Let us analyze the dynamics of the R^2 score. In both cases (the 70/30 and 80/20 training/test samples distributions) $R^2_{\max}$ decreases significantly slower than $R^2_{\min}$, which results in the rise of ΔR^2. This can be seen in the figures 4 – 7.

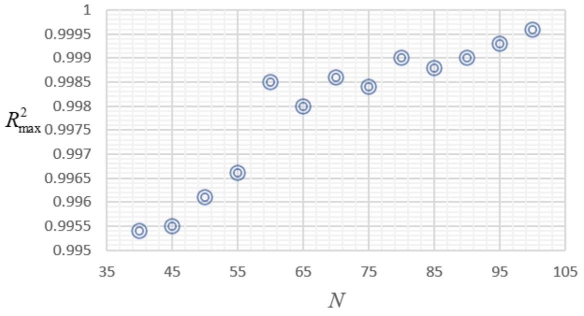

Fig. 4. Dynamics of $R^2_{\max}$ for the case of 70/30 training/test samples distribution.

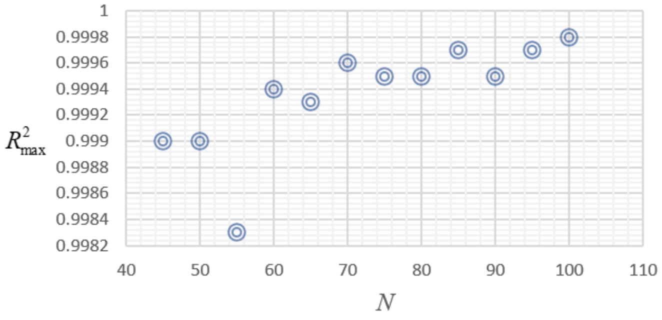

Fig. 5. Dynamics of $R^2_{\max}$ for the case of 80/20 training/test samples distribution.

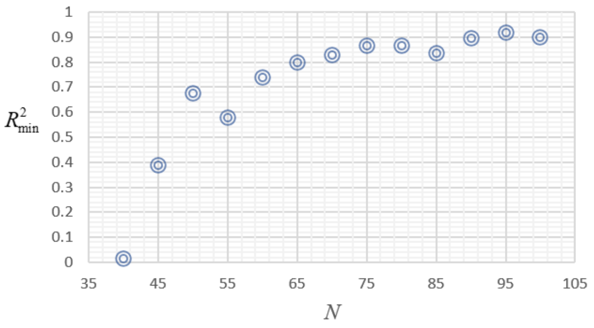

Fig. 6. Dynamics of $R^2_{\min}$ for the case of the 70/30 training/test samples distribution.

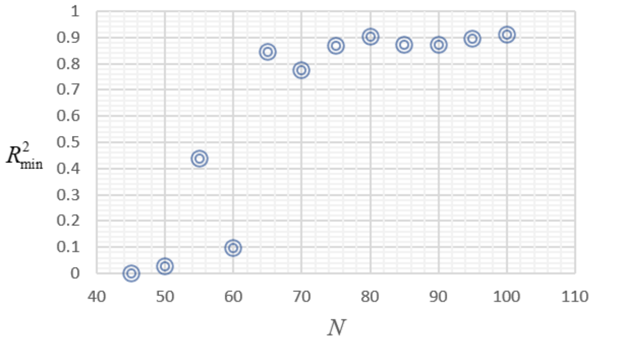

Fig. 7. Dynamics of $R^2_{\min}$ for the case of the 80/20 training/test samples distribution.

Considering the different dynamics of $R^2_{\max}$ and $R^2_{\min}$ decrease, let us analyze the dynamics of ΔR^2. It is shown in the figures 8 and 9. The values of ΔR^2 were obtained according to the expression (4).

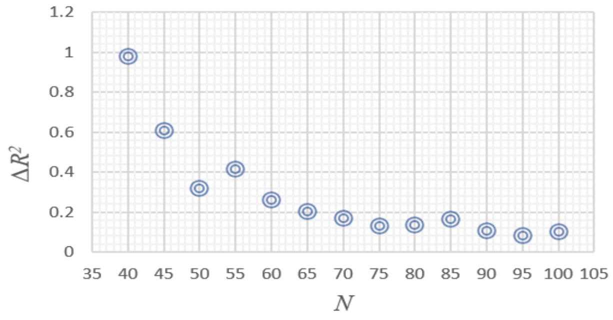

Fig. 8. Dynamics of ΔR^2 for the case of the 70/30 training/test samples distribution.

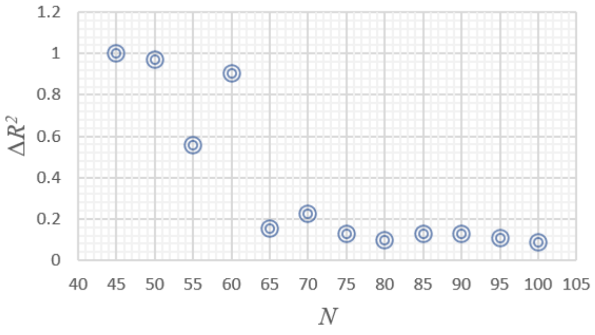

Fig. 9. Dynamics of ΔR^2 for the case of the 80/20 training/test samples distribution.

It is seen from the figures 8 and 9 that ΔR^2 tends to rise as the series size reduces. In both cases (the 70/30 and 80/20 training/test samples distributions) ΔR^2 tends to increase significantly after reaching 0.2.

By looking at the figures 8, 9 and the tables 1 and 2, it is possible to conclude that the first part of the hypothesis formulated above is confirmed because the R^2 score variation definitely increases with series size reduction. Considering the second part of the hypothesis, it is also confirmed because we do observe different values of the R^2 score obtained over the same size series, which can be seen in the tables 1 and 2. There we can find a minimum, maximum, and median values of R^2 obtained for the same N.

Let us consider the dynamics of R^2_{median} in order to point out a problem that may lead to incorrect model evaluation.

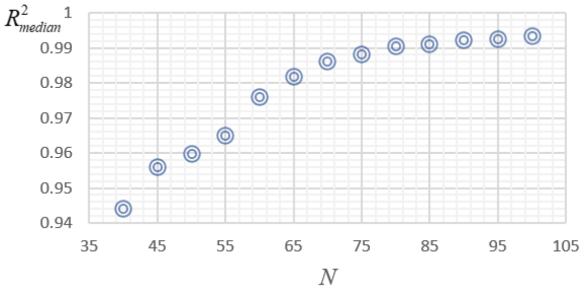

Fig. 10. Dynamics of R^2_{median} for the case of the 70/30 training/test samples distribution.

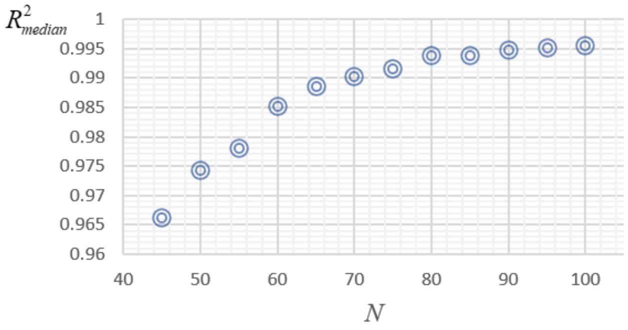

Fig. 11. Dynamics of R^2_{median} for the case of the 80/20 training/test samples distribution.

It is clearly seen from the figures 10 and 11 that the R^2_{median} stays rather high even in the case of large ΔR^2, which may lead to incorrect model evaluation. For example, let us examine the case of the 80/20 training/test samples distribution with N=50. In this case R^2_{median}= 0.9743 that seems to be a decent value, but $R^2_{\min}$ = 0.0268 that clearly indicates the underfit state. An existent solution is to use another metric, for example, MSE, MAE, ME along with the R^2 score to evaluate a model.

As the result of experiments carried out, it is possible to propose an alternative approach. In the figures 8 and 9 we can spot that the ΔR^2 tends to rise significantly after reaching 0.2. Thus, it is proposed not only to estimate the R^2 score value alone, but estimate it in the combination with the ΔR^2 value and the later must not exceed 0.2. Let us express this mathematically:

$$\Delta R^2 \leq 0.2 \wedge R^2 \rightarrow 1 \tag{5}$$

The expression (5) is the formulae of the proposed approach to increase the robustness and accuracy of a KNN regression model evaluation without the need to apply other metrics.

Conclusion

This work focused on researching the R^2 score variation caused by training series size changes and the randomness of selecting samples to training and test parts of series.

During the experiments, the hypothesis about the R^2 score variation increase with series size reduction was approved. Also, the hypothesis was approved in the part stating that random selection of samples to training and test parts of series resulted in R^2 score variation even for the same series size.

It is also shown that the KNN regression model evaluation based only on a single value of the R^2 score can be misleading.

The scientific novelty of the obtained results is that the current research paper proposes an alternative approach to KNN regression model evaluation. Precisely, it is proposed to combine the R^2 score metric with its variation ΔR^2 during KNN regression model evaluation. The value of ΔR^2 must satisfy the term (5). The proposed approach allows to increase the robustness and accuracy of an R^2 metric interpretation for a KNN regression model evaluation without the necessity of using other metrics suitable for regression models.

The practical significance of the obtained results is that the proposed approach does not require any supplementary metrics to be used during the KNN regression model evaluation – only the combination of the R^2 score and its variation ΔR^2, which simplifies software code and the process of metric interpretation.

Prospects for further research includes the question of how the obtained results can be generalized to other regression algorithms.

This work, as well, is aimed at rising a scientific discussion about application of the ΔR^2 along with the R^2 score metric for evaluation of regression models.

References

[1] *Sarkar, D., Bali, R., Sharma, T. (2018). Practical Machine Learning with Python. A Problem-Solver's Guide to Building Real-World Intelligent Systems. Apress Berkeley, CA, 545 p. DOI: 10.1007/978-1-4842-3207-1.*

[2] *Scikit-learn library web page. Sklearn.metrics.r2_score. Available at https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html*

[3] *Nakagawa S., Johnson, P., Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of The Royal Society Interface vol. 14(134), pp. 1–11. DOI: 10.1098/rsif.2017.0213.*

[4] Zhang, D. (2017) A Coefficient of Determination for Generalized Linear Models, The American Statistician, vol. 71:4, pp. 310–316, DOI: 10.1080/00031305.2016.1256839

[5] Gurubaran, K. et al. (2023). Machine Learning Approach for Soil Nutrient Prediction. 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2023, pp. 1-6. DOI: 10.1109/SILCON59133.2023.10405095.

[6] Gehlot, A., Sidana, N., Jawale, D., Jain, N., Singh, B.P., Singh,B. (2022). Technical analysis of crop production prediction using Machine Learning and Deep Learning Algorithms. 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-5. DOI: 10.1109/ICSES55317.2022.9914206.

[7] Tran, T. T. H., et al. (2022). Polygenic risk scores adaptation for Height in a Vietnamese population. 14th International Conference on Knowledge and Systems Engineering (KSE), Nha Trang, Vietnam, 2022, pp. 1-7. DOI: 10.1109/KSE56063.2022.9953620.

[8] Aulia, Y., Purnamasari, P.D., Zulkifli, F.Y. (2023). A Comparative Analysis of Machine Learning Algorithms for Predicting the Dimensions of Rectangular Microstrip Antennas. 2023 IEEE International Symposium On Antennas And Propagation (ISAP), Kuala Lumpur, Malaysia, 2023, pp. 1-2. DOI: 10.1109/ISAP57493.2023.10388517.

[9] Shashank, S., Gourisaria, M.K., Bilgaiyan, S. (2023). Weather Forecasting Based Shared Bike Demand Analysis using Machine Learning. 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6. DOI: 10.1109/ISCON57294.2023.10112160.

[10] Kumar, A., Mishra, S.K., Kejriwal, A. (2022). Prediction of Happiness Score of Countries by Considering Maximum Infection Rate of People by COVID-19 using Random Forest Algorithm. 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6. DOI: 10.1109/CONIT55038.2022.9847791.

[11] Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc., Sebastopol, CA., USA. 510 p. ISBN: 9781492032649.

[12] Pandas library web page via NumFOCUS Inc. Available at https://pandas.pydata.org/

[13] NumPy. The fundamental package for scientific computing with Python by NumPy team. Available at https://numpy.org/

[14] Scikit-learn library web page. Sklearn.neighbors.KNeighborsRegressor. Available at https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

[15] Matplotlib.pyplot by the Matplotlib development team. Available at https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html

ДИНАМІКА МЕТРИКИ R^2 ДЛЯ МОДЕЛІ РЕГРЕСІЇ KNN, НАВЧЕНОЇ НА ВИБІРКАХ РІЗНОГО РОЗМІРУ

Ю.Бабіч, Л. Глазунова, Т. Калініна, Я. Петрович

Державний університет інтелектуальних технологій і зв'язку, вул. Кузнечна, 1, 65023, Одеса, Україна

R2 або коефіцієнт детермінації часто використовується як метрика для оцінки регресійних моделей. Її можна застосовувати окремо, але зазвичай її поєднують з іншими метриками, щоб підвищити точність оцінки моделі. Метою роботи є дослідження динаміки метрики R2 регресійної моделі к-найближчих сусідів, навченої на серіях різного розміру, щоб запропонувати новий підхід для підвищення надійності та точності оцінки моделі, коли метрика R2 використовується самостійно, без застосування інших метрик. Як правило, значення метрики R2 вище 0,8 вважається прийнятним, тоді як оцінювана модель вважається достатньо точною. Однак такий спосіб інтерпретації оцінки R2 може призвести до невірної оцінки точності моделі, що і показано в запропонованій статті. Отримані результати чітко показують, що значення метрики R2 можуть суттєво відрізнятися в деяких випадках залежно від конкретних значень ознак, відібраних до тестової частини вибірки, яка використовується для оцінки моделі. Зазначене відхилення може сприяти завищенню точності моделі, що, у свою чергу, може призвести до некоректних результатів застосування моделі. Відомі методи підвищення точності оцінювання моделі передбачають використання інших метрик додатково. Натомість ця стаття зосереджена на підвищенні оцінки точності моделі без необхідності використання інших метрик. Динаміка метрики R2 досліджується за допомогою 25000 циклів навчання та оцінки регресійної моделі к-найближчих сусідів. Відбір значень до навчальної та тестової частин вибірки відбувався випадковим чином. Для всіх експериментів кількість сусідів є фіксованою та дорівнює значенню за замовчуванням n_neighbors=5 методу KNeighborsRegressor, наданого бібліотекою Sklearn. У роботі формулюється та підтверджується наступна гіпотеза про те, що варіація метрики R2, як очікується, збільшиться зі зменшенням розміру серії, і передбачається, що варіація буде спостерігатися для моделей, навчених на тій самій вибірці, через випадковість відбору навчальних/тестових значень. Проведені експерименти дозволили запропонувати альтернативний підхід, який не потребує додаткових метрик. Запропонований підхід передбачає застосування метрики R2 разом із її варіацією, яка не повинна перевищувати 0,2 для регресійної моделі к-найближчих сусідів.

Ключові слова: *розмір вибірки, метрика R^2, коефіцієнт детермінації, регресійна модель.*