International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Evaluation of Impact of Data Processing on Quality of Machine Learning Model

Solovei Olga[1], Solovei Bohdan[2]

Kyiv University of Civil Building and Architecture, Kyiv, Povitroflotsky Avenue, 31, 03680,

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Peremohy Ave, 37, Kyiv, 03056, Ukraine

Abstract: ***In the article is considered a question of how initial data processing impacts the quality of machine learning model. As a res***, Ukraine***ult of the research had been received models' quality score when data is processed by different methods: standardization and scaling. The summary specifies what data process method to be used depending on the chosen machine learning model. All calculations were performed by the functions from Phyton libraries.***

Keywords: data processing method, confusion matrix, F1-scope, mean absolute error, machine learning model

1. Introduction

The quality of the machine-learning model determines its ability to solve machine learning's tasks.

In [1], were identified effective methods for primary data processing in terms of obtaining the best values of the normal distribution parameters ($\mu = 0$ and $\sigma = 1$). As a result of the analysis of methods: standardization, scaling, logarithmic transformation and Box-Cox transformation were concluded that: standardization and scaling are the most appropriate, however, how the usage of those methods affects the quality of the machine learning model has remained unexplored.

Thus, there is a need to determine an effective data processing method, taking into consideration its impact on the qualitative assessment of the machine learning model.

2. Goal

The goal of current research is to evaluate effectiveness data processing methods: standardization on the basis of arithmetic mean and standard deviation (hereinafter "standardization_1"); standardization based on the median and interquartile range (hereinafter "standardization_2") and scaling taking into account their impact on the qualitative assessment of the machine learning model.

3. Main part

The research is performed on the set of data "Cervical cancer (Risk Factors) Data-Set" sourced from "UCI - Machine Learning Repository" [6]. The dataset was collected at 'University Hospital of Caracas' in Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients and the indicator when the patient was diagnosed with cervical cancer (1- "diagnosed"; 0 - "not diagnosed"). The data were collected to build a machine learning model to predict the risk of cervical cancer diagnosis.

In the scope of this analysis are included models: logistic regression, support vector classification; k-nearest neighbors; decision tree classifier and random forest. To assess the quality of the constructed models, were applied "F_1score" measure (1), which is a balanced assessment of the accuracy ("Precision") and completeness ("Recall") of the constructed model [3].

$$F_1\text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1),$$

In formula (1) for class where "cervical cancer is not diagnosed" the measures: "Precision" and "Recall" to be calculated per formula (2-3); for class, where "cervical cancer is diagnosed" – per formula (4-5).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2),$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Precision} = \frac{TN}{TN + FN} \quad (4),$$

$$\text{Recall} = \frac{TN}{TN + FP} \quad (5)$$

Formulas (2-5) include notions: TP – the number of correctly predicted cases when cervical cancer is not diagnosed; FP– the number of incorrectly predicted cases when cervical cancer is not diagnosed; FN- the number of incorrectly predicted cases when cervical cancer is diagnosed; TN - the number of correctly predicted cases when cervical cancer is diagnosed.

To assess the quality of built random forest model will be used mean absolute error, calculated by formula (6)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (6)$$

where

y_i - predicted value of risk of the cervical cancer diagnosis;

x_i - real value when cervical cancer diagnosis was diagnosed.

The research of the impact of data processing methods on

built machine learning model's quality to be conducted through steps:

1) Identify in the collected data a vector of data that will be predicted (further Y) and input data for prediction model (further X);
2) Divide sets X and Y into data to train model (further X_train; Y_train) and data to test model (further X_test; Y_test) in propotion 0.75/0.25;
3) Scale X_train and X_test sets of data (further X_train_scaled and X_test_scaled) by scaling method. Identify values of the normal distribution parameters;
4) Build logistic regression model based on X_train_scaled; Y_train;
5) Identify a vector of predicted values (further Y_pred) using data for model testing - X_test_ scaled;
6) Assess model's quality;
7) Repeat steps 2-6 when data is process by "standardization_1" and "standardization_2";
8) Repeat steps 2-6 for models: support vector classification; k-nearest neighbors; decision tree classifier and random forest;
9) Compare received models' quality figures for each method of data processing.

The results steps 1-8 execution is proposed in tables 1-5.

Table 1: Quality assessment when data is process by scaling method

Model	Confusion matrix		F_1score
logistic regression	TP=197	FP=2	0.99
	FN=3	TN=13	0.84
support vector classification	TP=197	FP=2	0.98
	FN=4	TN=12	0.8
k-nearest neighbors	TP=194	FP=5	0.97
	FN=4	TN=12	0.73
decision tree classifier	TP=193	FP=6	0.98
	FN=1	TN=15	0.81

Table 2: Quality assessment when data is process by "standardization_1" method

Model	Confusion matrix		F1score
logistic regression	TP=193	FP=6	0.98
	FN=1	TN=15	0.81
support vector classification	TP=195	FP=5	0.97
	FN=6	TN=10	0.65
k-nearest neighbors	TP=194	FP=5	0.97
	FN=7	TN=9	0.60
decision tree classifier	TP=193	FP=6	0.98
	FN=1	TN=15	0.81

Table 3: Quality assessment when data is process by "standardization_2" method

Model	Confusion matrix		F_1score
logistic regression	TP=197	FP=2	0.99
	FN=3	TN=13	0.84
support vector classification	TP=194	FP=5	0.97
	FN=5	TN=11	0.69
k-nearest neighbors	TP=194	FP=5	0.97
	FN=5	TN=11	0.69
decision tree classifier	TP=193	FP=6	0.98
	FN=1	TN=15	0.81

To estimate the impact of data processing on machine learning model's quality collect received values of F_1score measure and mean absolute error in table 4. Received values of normal distribution parameters are presented in table 5.

Values of F_1score measure

Table 4: Values of F_1score measure and mean absolute error

Data process method	scaling	"standardization_1"	"standardization_2"
Machine learning model	F1-score		
logistic regression	0.84	0.81	0.84
support vector classification	0.8	0.65	0.69
k-nearest neighbors	0.73	0.6	0.69
decision tree classifier	0.81	0.81	0.81
	MAE		
Randomforest	0.059	0.059	0.059

Table 5: Values of the normal distribution parameters after data is scaled

Data process method	μ	σ
scaling	0.05	0.18
"standardization_1"	0	1
"standardization_2"	0.05	0.32

4. Conclusion

The figures from table 5 obviously presents that the best normal distribution parameters had been received for method "standardization_1", however, for that methods are received the worst value of F_1score measure (referring to table 4), which means, method "standardization_1" has the most negative impact on machine model's quality among methods included in current research.

Slightly worse distribution parameters correspond to the method "standardization_2" but it's negative impact on quality assessment for models: support vector classification, k-nearest neighbors is also visible.

The scaling method has the least negative impact on the models' quality assessments while the values of the normal distribution parameters are inferior to the corresponding parameters received when data processed by methods "standardization_1" and "standardization_2".

Important to notice, that data processing methods included in current research have no impact on quality of the models: decision tree classifier and Random forest.

So, based on figures form table4 and table 5 can be concluded the following: when building machine learning models: decision tree classifier and Random forest the choice of method "standardization_1" will guarantee the best normal distribution parameters.

When logistic regression is used then method "standardization_1" can be preferred but for models: support vector classification and k-nearest neighbors, it is necessary to use the scaling method to prevent deterioration of the model results.

References

[1] Solovei O., Solovei B., Analysis of methods of data

processing of construction objects for machine learning models. // VII International Scientific and Practical Conference "Management of Technology Development" - Kyiv March 25 - 26, 2020. KNUBA, - Kyiv: 2020.-P.87-88
[2] Ralph Winters. Practical Predictive Analytics, June 2017, Publisher: Packt Publishing, ISBN: 9781785886188
[3] Giuseppe Bonaccorso, Machine Learning Algorithms - Second Edition, 2018, Publisher: Packt Publishing
[4] Sarah Guido, Andreas C. Müller. Introduction to machine learning with Python, 2016, O'Reilly Media, Inc.
[5] BOWERMAN, B.L., O'CONNELL, R.T., and KOEHLER, A.B. (2004) Forecasting, time series and regression: an applied approach, Thomson Brooks/Cole: Belmont, CA.
[6] Cervical cancer (Risk Factors) Data-Set

Author Profile

Solovei Olga – received PhD degree in Kyiv University of Civil Building and Architecture in 2013. Since year 2018 is working with data science topics. Takes position as Senior business analyst at Luxoft Ukraine and professor assistant of the department of applied mathematics in Kyiv University of Civil Building and Architecture.

Solovei Bohdan – received B.S. degree in National Technical University of Ukraine in 2020. Since year 2018 is working with data science topics in Python