COMPARATIVE ANALYSIS OF DIFFERENT REGRESSION METHODS FOR PREDICTING FIELD EXPLORATION DATA FROM A LEAD-ZINC DEPOSIT

Kremena Arsova-Borisova

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; E-mail: kremena.arsova@mgu.bg

ABSTRACT. A comparative analysis of different regression methods for predicting a useful indicator was performed on data from an operational study. Within the framework of the study, real data from an exploitation survey of a lead-zinc deposit were investigated by means the built-in Predict function of the Wolfram Mathematica v.13.

Key words: regression analysis, machine learning

СРАВНИТЕЛЕН АНАЛИЗ НА РАЗЛИЧНИ РЕГРЕСИОННИ МЕТОДИ ЗА ПРОГНОЗА ВЪРХУ ДАННИ ОТ ЕКСПЛОАТАЦИОННО ПРОУЧВАНЕ НА ОЛОВНО-ЦИНКОВО НАХОДИЩЕ

Кремена Арсова-Борисова

Минно-геоложки университет „Св. Иван Рилски“, 1700 София

РЕЗЮМЕ: Върху данни от експлоатационно проучване е извършен сравнителен анализ на различни регресионни методи за прогнозиране на полезен показател. В рамките на конкретното изследване са изследвани реални данни от експлоатационно проучване на оловно-цинково находище посредством вградена функция Predict на продукта Wolfram Mathematica v.13.

Ключови думи: регресионен анализ, машинно самообучение

Introduction

The continuous and increasing amount of data generated from various sources necessitates the need for effective methods for processing, analyzing and extracting additional information. In recent years, a lot of advancements have been done in the field of the so-called machine learning. Data mining and machine learning methods have great potential for data integration. They are widely used in various scientific and engineering fields, including the mining industry. Those methods can be used to solve various tasks such as classification, clustering, prediction, discovering functional dependencies, discovering patterns in large data sets, etc. The end goal for the mining enterprises is the prediction of the quality indicators in the ore and their metal content (Hristov V., St. Topalov, 2012). “Machine learning is becoming an appealing tool in various fields of earth sciences, especially in resources estimation” (Caté, Antoine & Perozzi, Lorenzo & Gloaguen, Erwan & Blouin, Martin, 2017).

The core concept in the field of machine learning is regression analysis. It is a statistical method for modelling relationships between dependent variables (targets) and independent variables (predictors). Information about the nature and form of a given dependence can be obtained with the help of regression analysis. It is used to find trends in data variation and aid predict values of dependent variables for new or missing values of independent variables.

In the present article, a comparative analysis of seven regression methods is made. They ate included in the built-in predict function (Predict) by the system of mathematical calculations and analysis Wolfram Mathematica v13.

In (Topalov St., V. Hristov, 2019) a similar comparative analysis of some of the methods was used to predict copper content data in ores, with the (Recipes module) of SoftStat STATISTICA 12.

Experimental Framework

The dataset was obtained from the operational study of spent exploitation blocks on horizon 540 (block №7, block №9 and block №11), horizon 590 (block №7, block №9 and block №11) and horizon 640 (block №11, block №13 and block №15 and block №17) of the Varba deposit.

The data, totaling 339, are aggregated by horizons. They are defined by the coordinates X, Y and contain information about the content of the main quality indicators - lead and zinc.

The distribution of Pb and Zn is given in Figure 1.

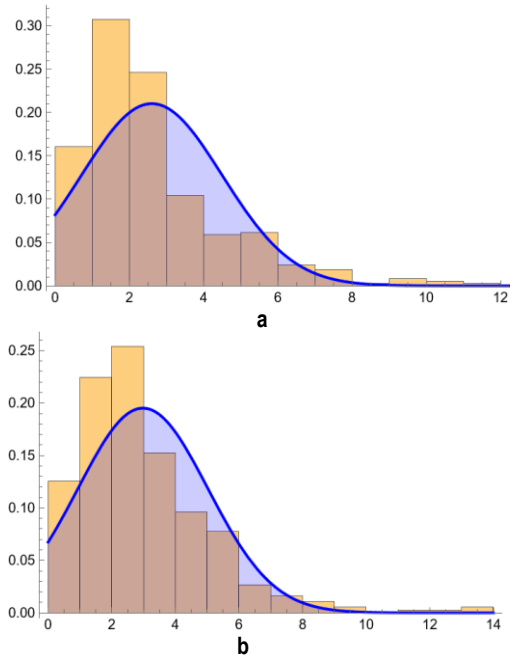

Fig 1. Distribution of useful component: lead (a) and zinc (b)

The built-in Predict function of the Wolfram Mathematica v. 13 software product was used to conduct the analyses. The methods included in the Predict function are:

Decision Tree (DT) – a supervised learning algorithm that can be used to solve both classification and regression problems. It can solve problems for both categorized and numerical data. When solving a regression problem, a tree structure is used in which each internal node represents the "test" for an attribute, each branch represents the result of the test, and each leaf represents the final solution or a result.

Gradient Boosted Trees (GBT) - used for both classification and prediction tasks. Gradient Boosted Tree works by building consistently simpler (weaker) prediction models, where each model tries to predict the error left by the previous model. These models are known as "weak students" because they are simple forecasting rules that perform slightly better than arbitrary ones.

Linear regression (LR) - a statistical regression method used for predictive analysis. Linear regression indicates the linear relationship between the independent variable and the dependent variable, which is why it is called linear regression. If there is only one input (independent) variable, then such a linear regression is called simple linear regression. And if there is more than one input variable, then such a linear regression is called multiple linear regression.

K-Nearest Neighbours (KNN) – a simple and versatile algorithm that stores all available cases and predicts the numerical target based on a similarity measure (e.g. distance functions). When a new item appears, it is compared to those that are already classified and searched for those closest to it and against the same measure. In this method, the right solutions are not necessarily the exact ones, but those that are the best possible.

Neural Network (NN) - a simplified model of neurons connected in the human brain. They perform parallel processing of information, enabling them to solve complex tasks with a small number of steps.

Random Forest (RF) – used to solve regression and classification tasks. It is an ensemble learning method that combines multiple decision trees and predicts the final result based on the average of each tree's result. Ensemble learning is a process in which multiple machine learning models are generated and combined to solve a particular problem.

Gaussian Process (GP) – a generalized supervised learning method designed for solving regression and probabilistic classification problems.

Results

The data on the values of the two useful indicators – lead and zinc are organized in tabular form in MS Excel spreadsheet. Three files were created for the three working horizons respectively – 540 (68 samples), 590 (97 samples) and 640 (174 samples). In separate tables, for each horizon, the lead and zinc content data from the exploitation study are contained, spatially determined by their coordinates (x, y) in a conditional coordinate system. The files were then exported to the Wolfram Mathematica v.13 environment.

Regression analysis is carried out separately for the two indicators using the above seven methods for each of the three horizons, with the content of the useful component being the dependent variable and the coordinates x and y is the independent. The results are shown in tables, for each horizon and its useful indicator. Each table contains: mean ($\bar{x}$) and variance (σ^2) of the original data; correlation coefficient (PCC) and the error (SE) of the predicted data for each method; mean and variance of forecasts.

Table 1 represents the results for Horizon 540 and the useful lead component. From the correlation coefficient values, it can be observed that the best result gives the method of Gradient Boosted Trees (GBT). Actual versus predicted data is shown in Figure 2. The predicted values obtained by the DT method for lead in this horizon are unsuitable for further analysis

Table 1. *Horizon 540, lead (Pb)*

$\bar{x}$ Pb, %	σ^2 Pb, %	Method	PCC	SE	Pred. $\bar{x}$ Pb, %	Pred. σ^2 Pb, %
2.57	5.29	DT	-	-	-	-
		GBT	**0.73**	**0.14**	**2.59**	**1.40**
		LR	0.20	0.06	2.57	0.21
		KNN	0.55	0.13	2.66	1.13
		NN	0.21	0.06	1.93	0.22
		RF	0.57	0.15	2.34	1.58
		GP	0.56	0.19	2.41	2.49

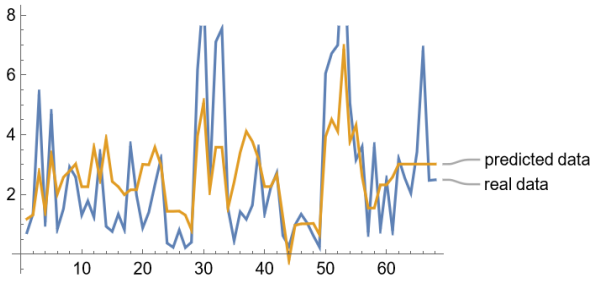

Fig 2. Actual versus predicted GBT method data for horizon 540 and lead (Pb) useful component

Table 2 represents the results for Horizon 540 and a useful zinc component. Actual versus predicted data is shown in Figure 3. Here the Gaussian Process (GP) method is best represented.

Table 2. *Horizon 540, zinc (Zn)*

$\bar{x}$ Zn, %	σ^2 Zn, %	Method	PCC	SE	Pred. $\bar{x}$ Zn, %	Pred. σ^2 Zn, %
2.4	4.18	DT	0.32	0.09	2.45	0.57
		GBT	0.68	0.02	2.41	0.02
		LR	0.35	0.09	2.40	0.50
		KNN	0.60	0.13	2.47	1.09
		NN	0.41	0.11	2.47	0.79
		RF	0.59	0.13	2.09	1.16
		GP	**0.87**	**0.19**	**2.41**	**2.49**

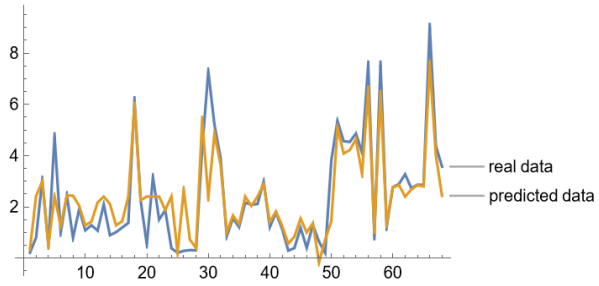

Fig 3. Actual versus predicted GP method data for horizon 540 and zinc (Zn) useful component

Table 3 presents the results for Horizon 590 and a useful lead component. Here can be observed that two methods weld the best results– Random Forest (RF) and Gaussian Process (GP). Actual versus predicted data is shown in Figure 4 (RF method) and Figure 5 (GP method). The predicted values obtained by the DT method for lead in this horizon are unsuitable for further analysis.

Table 3. *Horizon 590, lead (Pb)*

$\bar{x}$ Pb, %	σ^2 Pb, %	Method	PCC	SE	Pred. $\bar{x}$ Pb, %	Pred. σ^2 Pb, %
2.36	2.31	DT	-	-	-	-
		GBT	0.69	0.04	2.36	0.14
		LR	0.28	0.04	2.36	0.17
		KNN	0.58	0.08	2.33	0.69
		NN	0.33	0.05	2.29	0.24
		RF	**0.70**	**0.10**	**2.33**	**0.89**
		GP	**0.70**	**0.11**	**2.36**	**1.18**

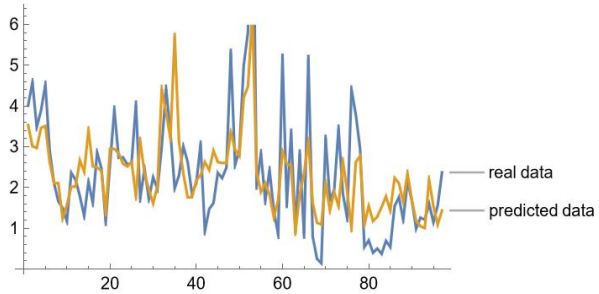

Fig 4. Actual versus predicted RF method data for horizon 590 and lead (Pb) useful component

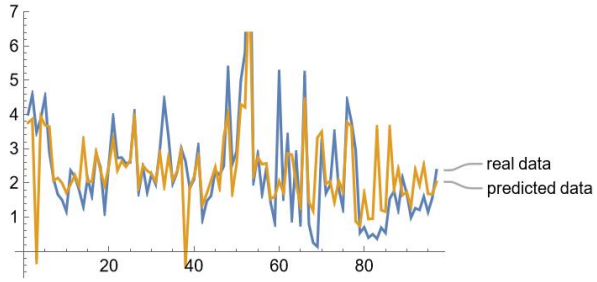

Fig 5. Actual versus predicted GP method data for horizon 590 and lead (Pb) useful component

Table 4 shows the results for Horizon 590 and useful zinc component. Here can be noted that the best result is given by Gaussian Process (GP). Actual versus predicted data is shown in Figure 6 (GP method).

Table 4. *Horizon 590, zinc (Zn)*

$\bar{x}$ Zn, %	σ^2 Zn, %	Method	PCC	SE	Pred. $\bar{x}$ Zn, %	Pred. σ^2 Zn, %
3.01	3.41	DT	0.58	0.11	3.01	1.13
		GBT	0.74	0.12	2.96	1.32
		LR	0.37	0.07	3.01	0.48
		KNN	0.58	0.10	3.00	0.97
		NN	0.26	0.05	2.29	0.24
		RF	0.70	0.10	2.94	0.99
		GP	**0.82**	**0.10**	**3.05**	**0.98**

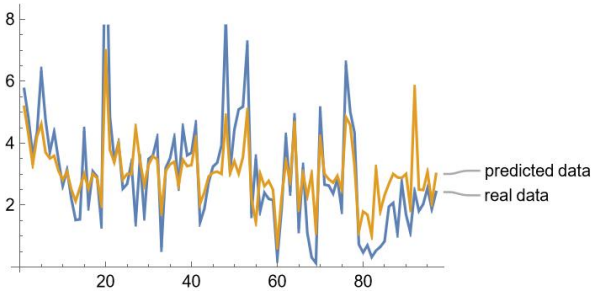

Fig 6. Actual versus predicted GP method data for horizon 590 and zinc (Zn) useful component

Table 5 presents the results for Horizon 640 and a useful lead component. The best result is given by the Gaussian Process method (GP). Actual versus predicted data is shown in Figure 7.

Table 5. *Horizon 640, lead (Pb)*

$\bar{x}$ Pb, %	σ^2 Pb, %	Method	PCC	SE	Pred. $\bar{x}$ Pb, %	Pred σ^2 Pb, %
2.83	3.96	DT	0.51	0.09	2.82	1.26
		GBT	0.74	0.09	2.82	1.40
		LR	0.31	0.05	2.83	0.39
		KNN	0.63	0.09	2.86	1.44
		NN	0.36	0.05	2.80	0.43
		RF	0.70	0.10	2.92	1.88
		GP	**0.88**	**0.11**	**2.77**	**2.10**

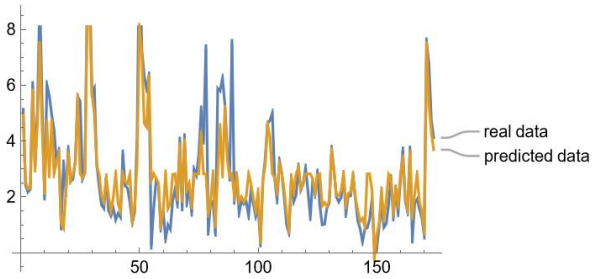

Fig 7. Actual versus predicted GP method data for horizon 640 and lead (Pb) useful component

Table 6 shows the results for Horizon 640 and a useful zinc component. The best results are given by the methods of Gradient Boosted Trees (GBT) and Gaussian Process (GP). Actual versus predicted data is shown in Figure 8.

Table 6. *Horizon 640, zinc (Zn)*

$\bar{x}$ Zn, %	σ^2 Zn, %	Method	PCC	SE	Pred. $\bar{x}$ Zn, %	Pred. σ^2 Zn, %
3.24	4.72	DT	0.54	0.12	3.29	2.41
		GBT	**0.89**	**0.12**	**3.21**	**2.66**
		LR	0.36	0.06	3.24	0.62
		KNN	0.68	0.11	3.26	1.95
		NN	0.31	0.05	2.80	0.43
		RF	0.73	0.11	3.37	2.24
		GP	0.75	0.07	3.22	1.01

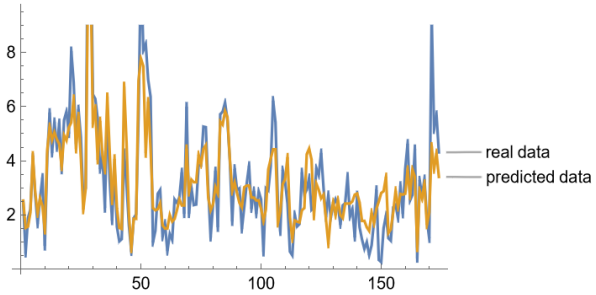

Fig 8. Actual versus predicted GBT method data for horizon 640 and zinc (Zn) useful component

Conclusions

As a result of the seven comparative analyses, the rating of the different methods based on correlation coefficients is as follows:

Table 7. *Rating of methods*

Method	Place in general	Place/lead	Place/zinc
DT	6	7	5
GBT	**2**	**1**	**2**
LR	7	6	6
KNN	4	4	4
NN	5	5	7
RF	3	3	3
GP	**1**	**2**	**1**

It can be seen that the methods GP (Gaussian Process) and GBT (Gradient Boosted Trees) give the best quality result both in the general case and separately for the two useful components. It is possible that with a more precise adjustment of the methods, this ranking could change, but not significantly.

References

Caté, Antoine & Perozzi, Lorenzo & Gloaguen, Erwan & Blouin, Martin. 2017. Machine learning as a tool for geologists. - *The Leading Edge. 36.* pp. 215-219. 10.1190/tle36030215.1.

Hristov, V., St. Topalov. 2012. Varianti na prognozirane s nevronni mrezhi pri podzemen dobiv na olovno – cinkova ruda. – Treta nacionalna nauchno–tehnicheska konferencija s mezhdunarodno uchastie „Tehnologii i praktiki pri podzemnija dobiv v minnoto stroitelstvo”, Devin, pp. 26 – 31. (in Bulgarian)

Topalov St., V. Hristov. 2019. Data mining of open pit mine-geological data. – *Proceedings of the XV International Conference of the Open and Underwater Mining of Minerals, Varna, Bulgaria,* 3-7 June, ISSN 2535-0854, pp. 261-267