Features of machine learning in the study of the main factors of development of countries of the world

Ludmila Borisova^{1,}, Galina Zhukova¹, Anna Kuznetsova², and Julie Martin³*

¹Financial University under the Government of the RF, 49 Leningrasky prosp., Moscow, 125993, Russia

²Institute of Biochemical Physics (IBCP), Russian Academy of Sciences (RAS), 4 Kosygina st., Moscow, 119334, Russia

³Health Service Executive, Steeven's Lane, Dublin 8, D08 W2A8 Republic of Ireland

Abstract. The paper analyzes the socio-economic and demographic indicators of life expectancy in the countries of the world. Methods of regression analysis and machine learning are used. Statistically significant indicators that affect life expectancy around the world have been identified. When analyzing the data using machine learning methods, 13 of the 14 analyzed indicators were statistically significant. Significant indicators, in addition to those selected in the regression analysis, were 3: the under-five infant mortality rate (per 1,000 live births), the Net Barter Terms of Trade Index (2000 = 100), and Imports of goods and services (in % of GDP) (in the regression analysis, only the infant death rate was significant). In addition, it should be noted that there is a significant decrease in the under-five infant mortality rate (per 1,000 live births) for the EU, CIS and South-East Asian countries compared to the border set in the study for all countries: 4.65 vs. 34.9, a decrease in the birth rate from 2.785 to 1.85, a sharp increase in exports of goods and services: from 23.17 to 80.59, a halving in imports of goods and services, a drop in population growth from 2.105 to 0.85. The performed statistical analysis strongly supports the use of machine learning methods in identifying statistically significant relationships between various indicators that characterize the development of countries, if there are gaps in the data.

The main indicators of the country's economic development are the Human Development Index (standard of living, literacy, education and longevity) and GDP per capita. The main drivers of economic growth and development are human capital and the innovations it generates. The paper [1] speaks about the importance of studying the phenomenon of life expectancy when modeling the human capital index, when solving the question of choosing a competence-based approach related to the development of cognitive abilities or a knowledge-based approach. In this paper, we analyzed the impact of the main socio-economic and demographic indicators presented on the World Bank website [2]. Selected

* Corresponding author: borisovalr@mail.ru

indicators for analysis: Life expectancy; High-tech exports; Net Barter Terms of trade Index (2000 = 100); Population growth (in % per year); Birth rate; Gini coefficient; Unemployment rate (%); GDP per capita (international dollars); Under-five child mortality rate (per 1,000 live births); Government spending on education, total (in % of GDP); Import of goods and services (in % of GDP); Export of goods and services (in % of GDP); Number of mobile subscribers (per 100 people); Inflation, GDP deflator (in % per year). These indicators are presented for most of the world countries. However, data on the Gini coefficient, high-tech exports, and government spending on education are not available for many countries around the world. Therefore, when we initially analyzed data for 186 countries of the world, representing all continents, we get an equation with a multiple correlation coefficient of 0.93, which speaks in favour of the indicators selected for further research, but the regression coefficients were statistically reliable only for 5 indicators: population growth (in % per year) (x_1), fertility rate (number of births per woman)(x_2), unemployment rate (%per year)(x_3), GDP per capita (international dollars)(x_4) and the under-5 infant mortality rate (per 1,000 live births) (x_5). Therefore, we build a regression based on these features. The regression equation has the form:

$$y=78.79+0,478x_1-1,365x_2-0,11x_3+0,00006x_4-0,18x_5. (1)$$

All coefficients are statistically significant (p-value < < 0.01). The multiple correlation coefficient is 0.94.

Figure 1 shows the coincidence of life expectancy with the theoretical curve, which is a multiple regression.

Fig. 1. Life expectancy for 186 countries and the theoretical curve calculated by equation (1) with the parameters specified in the text.

Next, we analyzed the data separately for different regions and found that for the most disadvantaged countries (countries in Africa, Central and South Americas), the most significant parameter determining life expectancy is the infant death rate. All the coefficients in the constructed regression equations were statistically significant (the analysis data is not presented), while for the EU countries, the indicators of the unemployment rate, the inflation rate and population growth (% compared to the previous year) were significant. For the rest of Europe and the former Soviet Union, GDP per capita was significant. After the regression analysis, we used machine learning methods to identify hidden links between the features, and to verify whether economic development

indicators such as imports of goods and services (in % of GDP), exports of goods and services (in % of GDP), the number of mobile subscribers (per 100 people), and the level of inflation (in % per year) have a weak impact on life in the analyzed countries. To do this, we used machine learning methods.

In solving the proposed problem, we used machine learning methods included in the Data Master Azforus computer complex, including both classical methods (adaptive and gradient boosting, decision trees, support vector methods, etc.), and methods of optimally reliable partitions (OVP) and statistically weighted syndromes (SWS) developed by the authors. Specific examples of the application of the described methods are considered and a comparative assessment of their capabilities for solving the project tasks is made. Previously, these methods were successfully applied in the work [3, 4].

In the first task, the training sample included all 188 countries. The target variable was "Life expectancy" with a division into two classes of 68 years. Total number of objects: 188. Number of objects of the first class: 59 (under the limit). Number of objects of the second class: 129 (above the limit). All indicators included in the training sample became significant when comparing the two studied classes according to the Wilcoxon-Mann-Uttney criterion. The only exception was unemployment.

The birth rate and population growth in the first class is twice as high as in the second. Government spending on education, combined (in % of GDP), is much higher in the second grade than in the first. GDP per capita (international dollars) is more than 6 times higher in the second class. The under-five infant mortality rate (per 1,000 live births) is 5.4 times higher in the first grade. High-tech exports are twice as large in the second class. Inflation, the GDP deflator (in % per year) is three times higher in the first class. The net barter terms of trade index is 24 points higher in the first class. On the contrary, imports of goods and services in the first class are 8% lower.

To clarify the boundaries between classes in the multidimensional feature space, we used the OVP method.

Table 1. The results of one-dimensional partitions for the studied classes obtained using the OVP method and the permutation test.

№	Indicator	Limit	Under the limit		Above the limit		F	p-value
			1 gr.	2 gr.	1 gr.	2 gr.		
1	Mortality rate, infant (per 1,000 live births)	34,90	3 (5,1 %)	125 (99,2 %)	56 (94,9 %)	1 (0,8 %)	166,087	0,0005
2	birth rate	2,785	5 (8,6 %)	115 (89,1 %)	53 (91,4 %)	14 (10,9 %)	112,24	0,0005
3	GDP per capita (international dollars)	5736	49 (84,5 %)	20 (15,5 %)	9 (15,5 %)	109 (84,5 %)	81,327	0,0005
4	Population growth (in % per year)	2,105	20 (33,9 %)	119 (92,2 %)	39 (66,1 %)	10 (7,8 %)	71,146	0,0005
5	Number of mobile subscribers (per 100 people)	90,5	41 (69,5 %)	15 (11,8 %)	18 (30,5 %)	112 (88,2 %)	63,351	0,0005
6	High-tech exports	4,05	29 (67,4 %)	36 (31,9 %)	14 (32,6 %)	77 (68,1 %)	16,122	0,0005
7	Exports of goods and services (in % of GDP)	23,17	22 (38,6 %)	11 (8,7 %)	35 (61,4 %)	116 (91,3 %)	23,823	0,0005
8	Gini Coefficient	39,90	19 (35,2 %)	72 (70,6 %)	35 (64,8 %)	30 (29,4 %)	18,091	0,0015
9	Inflation, GDP deflator (in % per year)	2,625	20 (36,4 %)	87 (68 %)	35 (63,6 %)	41 (32 %)	15,738	0,002
10	Government expenditure on education, total (% of GDP)	3,15	11 (42,3 %)	10 (11,1 %)	15 (57,7 %)	80 (88,9 %)	13,128	0,0065
11	Imports of goods and services (in % of GDP)	44,24	38 (70,4 %)	53 (42,1 %)	16 (29,6 %)	73 (57,9 %)	12,05	0,0105
12	Net Barter Terms of Trade Index (2000 = 100)	101,19	11 (19,6 %)	56 (44,8 %)	45 (80,4 %)	69 (55,2 %)	10,44	0,0325

As can be seen from Table 1, 7 indicators are significant with $p<0.0005$. Thus, the under-five infant mortality rate exceeds the value of 34.9 per 1,000 live births for 94.9 % of class 1 countries. The birth rate is below the 2.785 value for 89.1 % of the class 2 countries. GDP per capita is below the value of 5736 international dollars. for 84.5 % of class 1 countries. Population growth is below 2,105 % per year for 92.2 % of class 2 countries.

2.2 Positioning

Fig. 2. The X-axis is the under-five infant mortality rate (per 1,000 live births). On the Y-axis, GDP per capita (international dollars). In quadrant I, class 2 prevails – with a life expectancy of more than 68 years. In quadrant II, class 1 prevails - life expectancy is less than 68 years. The significance of the found pattern is $p<0.005$.

Figure 3 shows that the values of class 1 countries (with low life expectancy) are mainly in the upper right quadrant, i.e., with high population growth and a high Gini coefficient (the degree of stratification of society).

The ensemble was created using 5 machine learning methods: AdaBoost, GBM, DT, KNN, and SVM. The area under the ROC curve was $AUC=0.990$ (Fig. 4).

Recognition errors on the sliding control occurred in Class 1: Syria, Sao Tome and Principe, South Africa and Guyana. These countries were assigned to the second (prosperous in terms of life expectancy) class. This may indicate a positive trend in the standard of living in these countries. In the second class, East Timor and the Dominican Republic were mistakenly recognized. They were assigned to the 1st class, which can be regarded as a negative trend in the standard of living. Further, a comparative analysis was carried out for different groups of countries. Comparison of EU, CIS, and Southeast Asian countries by GDP per capita. The boundary of the division into classes is 19,334 international dollars (the division was carried out according to the median). The training

sample included 65 countries of the EU, CIS and South-East Asia. Class 1-34 countries with GDP per capita less than 19,334 international dollars, class 2-31 countries with GDP above this threshold.

Fig. 3. The X-axis is the under-five infant mortality rate (per 1,000 live births). On the Y-axis, GDP per capita (international dollars). The other three quadrants are dominated by Class 2, with a life expectancy of less than 68 years. The significance of the found pattern is $p < 0.005$.

Fig. 4. The ROC-curve on the ensemble of machine learning methods showed AUC=0.990.

When compared by the U-criterion, in addition to socio-demographic indicators, the following indicators were also significant in descending order of importance: Trade in goods (in % of GDP), the Index of the number of the poor, the Gini Coefficient, Inflation, the GDP deflator (in % per year), Public spending on education, total (in % of GDP), Exports of high technologies (% of exports of manufacturing products), the Index of Net Barter terms of trade (2000 = 100), Imports of goods and services (in % of GDP).

Fig. 5. On the X-axis – Life expectancy. On the Y-axis, the under-five infant mortality rate (per 1,000 live births). Quadrants II and III (right, green circles) are dominated by Class 2-with a GDP per capita of more than \$ 19,334 international dollars. The I quadrant is dominated by Class 1 (red crosses). The significance of the found pattern is $p < 0.005$.

The next task was to study the countries of North, Central and South Americas in relation to the same sociological, demographic and economic indicators in terms of the division by Life expectancy (Fig. 5). The limit of the division was 75 years old. The training sample included 35 countries. Number of first class objects: 18. Number of objects of the second class: 17.

Table.2. The results of one-dimensional partitions obtained using the OVP method and the permutation test of the American countries, divided by the indicator of life expectancy.

№	Indicator	Limit	Under the limit		Above the limit		p-value
			1 gr.	2 gr.	1 gr.	2 gr.	
1	Imports of goods and services (in % of GDP)	27,42	0 (0 %)	9 (52,9 %)	18 (100 %)	8 (47,1 %)	0,0025
2	Mortality rate, infant (per 1,000 live births)	17,75	8 (44,4 %)	16 (94,1 %)	10 (55,6 %)	1 (5,9 %)	0,0115

We see (Table 2) that the limit on the rate of child mortality under the age of five is twice lower (17.75) than it is in the task for the countries of the world (34.9). Imports of goods and services (in % of GDP) are also significantly lower than in Table 1 - 27.42

against 44.24% of GDP. Only 3 indicators were significant when using the OVP and SWS methods in this task: the under-five infant mortality rate (per 1,000 live births), the Net Barter Terms of Trade Index (2000 = 100), and Imports of goods and services (in % of GDP). Countries with high life expectancy show a Net Barter Terms of trade Index above 101.4. At the same time, their infant mortality rate is below 17.75. Conversely, countries in Class 1 have either low values of both indicators or high values.

When analyzing a sample of EU and CIS countries, the total number of objects was 42, class 1-below 78 years of life expectancy – 20 countries, class 2-above this age – 22 countries. Significant indicators were: GDP per capita with a breakdown of 22,862 international dollars; Inflation, GDP deflator (in % per year) - 2.7; Under-five child mortality rate (per 1,000 live births) - 4.65; Trade in goods (in % of GDP) - 72.5; Net Barter Terms of Trade Index (2000 = 100) - 94.5; Population growth (in % per year) - (-0.35).

You can compare these values with the countries of Africa and the former socialist republics. This is the last task that was solved in the framework of this study. 71 countries were included in the training sample. Number of first class objects: 30, Number of second class objects: 41. Significant indicators in the OVP analysis were the birth rate with a division boundary of 2.13; the under-five infant mortality rate (per 1,000 live births) - 11.1; Population growth (in % per year) - 0.3; Life expectancy-66.9; Exports of goods and services (in % of GDP) - 22.35; Imports of goods and services (in % of GDP) - 50.41; High-tech exports-1.5.

From the above data, we can see how large the gap in the values of indicators is for the EU, CIS and African countries.

Thus, the performed statistical analysis strongly supports the use of machine learning methods in identifying statistically significant relationships between various indicators that characterize the development of countries, if there are gaps in the data. In addition, despite the high life expectancy in many European countries, there is a very low annual population growth, sometimes even a decline, and a low number of births per woman compared to many countries in Asia, America and Africa. To predict the human capital index and the index of confidence in the future, psychological indicators are also needed for analysis.

References

1. Kamenev K. G., Kamenev N. G. Multidimensional metric analysis in the modeling of human capital. *Computer Research and Modeling*, vol. 12, no. 5, pp. 1223-1245 (2020).
2. <https://data.worldbank.org>.
3. L.R. Borisova, A.V. Kuznetsova, N.V. Sergeeva, O.V. Senko. *Computer Research and Modeling*, vol. 12, no. 1, pp. 201-215 (2020).
4. Kirilyuk I.L., Volynsky A.I., Kruglova M.S., Kuznetsova A.V., Rubinstein A.A., Sen'ko O.V. Empirical testing of institutional matrices theory by data mining. *Computer Research and Modeling*, 2015, vol. 7, no. 4, pp. 923-939.