Classical Machine Learning Methods in Economics Research: Macro and Micro Level Examples

VITALINA BABENKO
Department of International E-Commerce and Hotel&Restaurant Business
V.N. Karazin Kharkiv National University
Svobody sq., 4, Kharkiv, 61022
UKRAINE

ANDRIY PANCHYSHYN
Department of Economic Cybernetics
Ivan Franko National University of Lviv
Universytetska Street, 1, Lviv, 79000
UKRAINE

LARYSA ZOMCHAK
Department of Economic Cybernetics
Ivan Franko National University of Lviv
Universytetska Street, 1, Lviv, 79000
UKRAINE

MARYNA NEHREY
Department of Economic Cybernetics
National University of Life and Environmental Sciences of Ukraine
Heroiv Oborony Street, 15, Kyiv, 03041
UKRAINE

ZORIANA ARTYM-DROHOMYRETSKA
Department of Economic Cybernetics
Ivan Franko National University of Lviv
Universytetska Street, 1, Lviv, 79000
UKRAINE

TARAS LAHOTSKYI
Department of Economic Cybernetics
Ivan Franko National University of Lviv
Universytetska Street, 1, Lviv, 79000
UKRAINE

Abstract: - Paper reviews the classical methods of machine learning (supervised and unsupervised learning), gives examples of the application of different methods and discusses approaches that will be useful for empirical economics research (on data from Ukrainian firms, banks and official state statistics). The different sectors of economics are investigated: the multiple linear regression is used on macrolevel for macro production function of Ukraine specification; logistic regression is used in bank sector for credit risk management with the scoring model; k-means, hierarchic clustering and DBSCAN are used in regional level for regions of Ukraine grouping based on competitiveness; principal component analysis is used for firm's financial stability analysis. All models showed adequate simulation results according to the quality criteria of the models. So, the possibility of classic machine learning methods application for investigations of the processes and objects on different levels of economics (micro, mezzo and macro) is demonstrated in the article.

Key-Words: - Machine Learning, Economics, Regression, Classification, Clustering, Modeling.

1 Introduction

The rapid growth of data and the increase in computing power contribute to the increased interest in machine learning and the opportunities that are opened by machine learning methods. New algorithms under development focus on many aspects of the economy and generate new fields of research in the field of economics. The successful application of machine learning to researchers requires basic knowledge and understanding of critical concepts that were previously unfamiliar to economists.

In this article, we provide an overview of Machine Learning methods that, in our view, should be in the arsenal of scientist economist tools and be taught in master's programs in economics. This list is subjective and will change over time. We do not do a thorough analysis of all the methods, but we offer to get acquainted with the basic ideas and basic concepts of classical methods of machine learning.

Machine learning is the concept that a computer program can learn and adapt to new data without human interference [1]. Machine learning is a field of artificial intelligence that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy.

Machine learning includes the next methods classes: Supervised Learning, Unsupervised Learning, Reinforcement Learning, Ensemble methods, Neural Networks and Deep Learning. Classical machine learning methods (Supervised Learning and Unsupervised Learning) are presented in Table 1.

Table 1. Classical machine learning methods

Class	Subclass	Method
Supervised Learning	Regression	Linear regression
		Polynomial regression
		Ridge/lasso regression
	Classification	Logistic regression
		Decision trees
		Support vector machine
		Naïve Bayes
		k-nearest neighbor
Unsupervised Learning	Clustering	k-means
		Agglomerative
		Mean-shift
		Fuzzy C-means
		DBSCAN
	Rule Engine	Association rules
		Eclat
		Apriori
		FP-growth
	Dimensionality Reduction	Principal Component Analysis
		Partial Least Squares Regression
		Principal Component Regression
		T-distributed Stochastic Neighbor Embedding
		Singular Value Decomposition
		Mixture Discriminant Analysis
		Linear Discriminant Analysis

We discuss Supervised learning methods and give examples of their application in economic research, including linear regression to build a macroeconomic production function of Ukraine and logistic regression to build a scoring model. Next, we discuss methods of Unsupervised Learning: clustering methods (k-means, hierarchic clustering, DBSCAN) for the region of Ukraine clustering according to its competitiveness level, Association rules (apriori method) for consumer market basket analysis, Principal Component Analysis for firm financial stability analysis.

2 Problem Formulation

The machine learning methods applications in various scientific fields are becoming increasingly popular and economics as social science is not an exception. That's why it is interesting to investigate what is similar and what is different in application of machine learning methods at different levels of economic systems, which methods are suitable for different economic problems and so on. We propose to use supervised and unsupervised methods of classical machine learning for investigation for macroeconomic development of Ukraine, regional competitiveness of Ukraine, banks credit system effectiveness and firms financial stability.

3 Materials and Methods

3.1 Literature review

The application of machine learning methods in different sectors of the economy is considered in subsequent studies. Research [2] provides a discussion of the most popular techniques used in energy economics papers. In paper [3] is focusing on the benefits of the systematization of business processes in aviation by using Machine Learning methods. With Machine Learning methods investigated if hidden information can be used to macroeconomic prediction [4]. Paper [5] provides a discussion on recent developments in adapting Machine Learning methods for applications in education, particularly in economics. In [6] is focusing on the advantages and disadvantages of using supervised learning methods for econometric studies. Regression methods, including linear, ridge, and lasso used for economic time series forecasting [7, 8], for forecasting financial and macroeconomic variables [9], for corporate failure predicting [10], and so on. The decision tree method is used both individually and in comparison with other techniques: macroeconomic indicators prediction

[11], financial credit risk assessment [12], and transit service quality analysis [13]. Support vector machine method realized for credit scoring modeling [14]. The examples of using wavelet analysis [15] and Naïve Bayes methods in stock market modeling [16].

Unsupervised Learning methods are widely used in various fields of economic research. For example, in the research of the relationship between financial market structure and the real economy [17], macroeconomic time series [18], the regional competitiveness [19]. The rule engine methods used in researches [20]. The Principal Component Analysis (PCA) applied in macroeconomics research [21], in finance [22], in the analysis of food quality and safety [23].

3.2 Research methodology

Regression. Regression analysis is a method of scientific research that is closely related to correlation analysis, but it is not only used to determine the form and discover the relationships between the dependent and independent variables, but also allows to identify the analytical form of the model. Regression analysis is used to determine the analytical form of the relationship in which the measurement of a depended variable is due to the influence of one or more independent variables, and the set of all other factors that also affect the performance trait acquire fixed or average values.

A regression function is a function that describes the one-way stochastic dependence of one random variable on another or several other random variables. The purpose of the regression analysis is to evaluate the functional dependence of the conditional mean value of the performance trait on the factor traits.

The regression equation characterizes the law of relation between dependent and independent variables, not for individual elements of the population, but for the population as a whole. The aim of constructing a linear regression model is to find the smoothing line that "best" passes through a given set of points. The least square method is most often used to estimate unknown parameters of linear regression.

Classification. The prediction with the logistic regression gives outcome as the probability that can have only two values (ie, a dichotomy). The logistic regression procedure is rather similar to multiple linear regression, except that fact that the dependent variable is binomial. The result is the effect of each independent variable on the event coefficient. The forecast is based on the use of one or more predictors (numeric and categorical).

The binary logistic regression response format will be 1 or 0. Having found the parameters of logistic regression, we will have a certain plane, specific points of examples, depending on the unit of this or zeros, must be on different sides of the plane. Then, by testing the system on a test sample, it will be possible to tell which class it is likely to belong to. The probability will range from 0 to 1. This is a kind of multiple regression, general purpose of which is to analyze the relationship between several independent variables (also called regressors or predictors) and the dependent variable. Logistic regression is best suited when the output variable accepts only two values.

Logistic regression coefficients usually are estimated with the maximum likelihood estimation method.

There is no equivalent statistic for the R-squared in logistic regression. Logistic regression model estimates are estimates of maximum likelihood through an iterative process. They are not designed to minimize variation. However, several pseudo-R squares have been developed to evaluate the quality of logistics models. They are called "pseudo" R-squared because they look like R-squared in the sense that they have a similar scale, ranging from 0 to 1 (though some never reach 0 or 1) with higher values, which indicates better model fit, but they cannot be interpreted as interpreting an R-squared and different pseudo-R-squared can acquire very different meanings.

Clustering. The purpose of cluster analysis is to classify objects into relatively homogeneous groups, taking into account the quantitative indicators of the objects. A cluster is a set of homogeneous elements, and the main task of cluster analysis is to form such sets in multidimensional space. Due to its specificity, the cluster analysis is based on the mechanism of development of management decisions aimed at combining economic objects of different functional orientation. For example, clustering of industries or the formation of territorial zones with different levels of indicative features.

There are many methods of grouping objects into clusters. The most popular method of clustering is the k-means method. The input set is divided into *k* groups, while minimizing the function that defines distances as Sum of Squared Errors (SSE).Currently, there are many variations of this method that partially eliminate the disadvantages, among them: K-Medoids, K-Medians, K-Modes, K-means ++, Intelligent K-Means, Genetic K-Means.

Statistical processing often uses hierarchical clustering (also called hierarchical cluster analysis or HCA), which is a cluster analysis method that seeks to build a cluster hierarchy. The result of hierarchical algorithms is a dendogram (hierarchical tree). A dendrogram is a tree constructed on a matrix of proximity measures. The dendrogram allows to depict the interconnections between objects of a given set. To construct a similarity (difference) matrix, it is necessary to specify a measure of the distance between two clusters. The following sorting strategies are most commonly used: single linkage, complete linkage, pair-group method using arithmetic mean, UPGMA, WPGMA, pair-group method using the centroid average, Ward's method.

Hierarchical clustering is worse for clustering large volumes of data than the k-means method. The k-means method is more sensitive to noisy data than the hierarchical method. In k-means clustering, the algorithm starts from an arbitrary selection of starting points, so the results generated by running the algorithm multiple times may differ. At the same time, in the case of hierarchical clustering, the results can be reproduced.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm used in data analysis as one of the k-means replacements. This is a density-based clustering algorithm - if a given set of points is in some space, the algorithm groups together points that are closely spaced (points with many close neighbours), denoting as emissions points that are lonely in low-density regions (nearest neighbours of which are lying far). DBSCAN is good in separating high-density clusters from low-density clusters within dataset, but it does not work well with similar density clusters.

Dimensionality Reduction. The principal component method detects -components - factors that explain all the variance and correlations of the initial *k* random variables; the components are plotted in descending order, explaining the fraction of the total variance of the original values, which often allows us to limit ourselves to the first few components. The first principal component of F_1 determines the direction in the space of the source traits by which the set of objects (points) has the greatest variation. The second principal component of F_2 is constructed in such a way that its direction is orthogonal to the direction of F_1 and it explains as much of the residual variance as possible, etc., down to the *k*-th principal component of F_k. Since the selection of the main components is in descending order in terms of the proportion explained by the variance, the features included in the first major component with large coefficients, maximally affect the differentiation of the studied objects. Such conversion allows the information to be reduced by discarding coordinates corresponding to the directions with minimal dispersion.

Let the activity of economic objects be characterized by a set of factors x_{ki}, where *i* is the number of the factor (i = 1,2,3, ..., n), *k* is the number of the economic object (k = 1, 2,3, ..., m), *n* is the number of factors, m is the number of economic objects. The values of each factor for different economic objects form the vector $x_i = \{x_{1i}, x_{2i}, \ldots, x_{mi}\}$ The space of economic object factors can be represented as a matrix of output factors *X*, where each column of the matrix contains the values of one factor (*i*) for different states of economic objects, and each row contains the values of all factors of the same state of the economic object. Thus, the state space of economic objects will be described as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1i} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2i} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{k1} & x_{k2} & \ldots & x_{ki} & \ldots & x_{kn} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & x_{mi} & \ldots & x_{mn} \end{bmatrix}$$

The centered and normalized factor space will be denoted by the matrix Z. Based on this matrix, we construct a matrix of even correlation coefficients R and estimate the eigenvalues and eigenvectors of the correlation matrix R. The eigenvalues of characterize the contributions of the corresponding principal components to the total (total) variance of the input features $x_1, x_2, \ldots, x_m$; the first component has the greatest effect on the total variance, the second one has the least, etc., the latter has the least effect on the total variance of all input features. The matrix of principal components is formed from the eigenvectors corresponding to the largest eigenvalues. The matrix of principal components allows to form new factors in the form of sum of initial normalized-centered factors. Thus, the main factor is a linear combination of the output factors of economic objects, and the main component is a set of weighting factors on the basis of which this combination is formed. The first principal component makes it possible to form the principal factor which, among other linear combinations (principal factors), has the largest variance. In this case, the *i*-th main factor is a linear combination of initial factors of the economic object, which is not correlated with *k-1* previous main factors, and among the other main factors has the largest variance.

4 Results and Discussion

4.1 Regression – Modeling of macroeconomic production function of Ukraine

The production function is known as an economic and statistical model of the production process in an economic system and expresses a consistent regular quantitative relationship between the volume indicators of resources and output. The production level in the general economy or corporate environment is described by the production function.

Based on the preliminary analysis, it was decided to use an econometric dependence, which looks like:

$$Y=\alpha_0 L^{\alpha 1} K^{\alpha 2}, \quad (1)$$

where Y - GDP, mln UAH; L - number of the working population aged 15-70 years, thousand people; K - capital investment, mln UAN.

Statistical information from the State Statistics Service of Ukraine for the period from 1996 to 2018 was used to build the econometric model of the macroeconomic production function of Ukraine [24].

After raising the parameter α_0 to the exponent the resulting model will look like

$$Y=5\ 246\ 882\ 879.6L^{-1.8731}K^{0.8202}. \quad (2)$$

Multiple coefficient of determination $R^2 = 0.9887$, and thus under the influence of changes in the number of the working population and the value of the capital investment, 98.87% of the variation in GDP is explained. Accordingly, the multiple correlation coefficient R = 0.9943. This value indicates that the relationship between the value of GDP and the number of able-bodied population and the value of capital is very close. The Fisher F-test. F-statistic = 876.17. A critical value of Fisher's F-criterion with $\nu 1 = 2$ and $\nu 2 = 19$, and 0.01 significance level, was found to be 5.33. Since Fem> Fcrit, the constructed model can be considered adequate. MAPE (Mean Abs. Percent Error) is used to calculate the quality of the forecast. The value of MAPE is 10.52%, which indicates a rather high quality of the forecast.

According to research by economists, innovation has a significant impact on economic growth, so it is advisable to investigate the dependence of GDP not only on the size of the working population and capital investment, but also on the volume of innovation activity.

An innovation model will look like:

$$Y=\alpha_0 L^{\alpha 1} K^{\alpha 2} I^{\alpha 3}, \quad (3)$$

where Y – GDP, mln UAH; L – number of the working population aged 15-70 years, thousand people; K – capital investment, mln UAH; I – the amount of expenditure on innovation, mln UAH. The model of macroeconomic production function with innovation will look like:

$$Y=2\ 094\ 082\ 035.23L^{-1.7835}K^{0.8945} I^{-0.0957}. \quad (4)$$

The model fairly describes the relationship between GDP and the working population, the size of capital investment, and the volume of innovation activity. Under the influence of changes in the number of the working population, the value of the capital investment, and the volume of innovations, 98.87% of the variation in GDP is explained. The MAPE value for the forecast based on the model with innovation is 10.42%, which indicates a slightly higher quality of the forecast than the model without taking into account the innovation.

Therefore, the built-in macroeconomic function based on innovation produces rather poor results, the forecast based on this model has a margin of error of 10%, which is acceptable in economic research.

4.2. Classification – Modeling of credit risk management.

The logistic regression method is used to build scoring models. Scoring models help reduce the risk of loan portfolios. The modeling is carried out in several stages: the selection of variables, analysis of descriptive statistics, estimation of model parameters, selection of an adequate model [25].

Independent variable models: age of the borrower; sex of the borrower; the marital status of the borrower; the number of dependents; monthly income; the borrower's residency at the last address; type of apartment; guarantors; the number of loans returned; and so on. The dependent variable of the model is the presence/absence from the client of debts on the previously taken loans (Creditability).

Checking the model's forecasting capabilities showed that the sensitivity of the model is 91%, specificity - 80%, which means that 91% of trustworthy borrowers will be identified by the model received, and 9% of bad borrowers will get credit.

To evaluate the quality of the model classification, a ROC curve was constructed, which shows the dependence of the number of correctly classified positive results on the number of incorrectly classified negative consequences (Fig. 1). The area under the curve is 0.90, which indicates the high predictive power and reliability of the model.

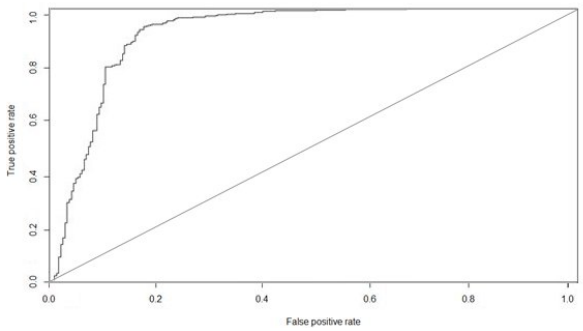

Fig.1. Scoring models ROC curve

For a more detailed analysis of the model, a classification matrix has been constructed (Table 2). It is shown that on the training set, the model more often refused to give credit to "good" borrowers than to give credit to "bad" ones. Approval Rate is 65%, Bad Rate is 5.55%.

Table 2. Classification matrix

	Actual		
Prediction	0	1	Sum
0	283	36	319
1	68	613	681
Sum	351	649	1000

In practice, it is possible to lower the cut-off threshold and make the model more likely to make a positive decision. The percentage of bounce rates will decrease but credit risk will increase accordingly. Therefore, the choice of a cut-off point depends on the set goals: to reduce the share of "bad" loans or increase the loan portfolio, more often making a positive decision on the client.

4.3. Clustering – Regions of Ukraine clustering by competitiveness

The 22 regions of Ukraine statistics for 2018 was taken (except Donetsk and Lugansk regions because of lack of statistics) [24]; economic indicators are used: the disposable income, gross regional product, industrial production, agricultural indices, the volume of industrial production sold, employment rate, natural population growth, waste generation, average monthly salary, and environmental protection costs. The 3 main directions of regional development are investigated: socio-economic economic, and ecological. With the correlation matrix closely correlated indices were eliminated. The final set of factors included 7 ones:

- economic: agricultural production index; industrial production index; GRP per person;
- socio-economic: average monthly salary per 1 person, the employment rate of the working-age population; the natural movement of the population;
- environmental: waste generation.

Before the k-means method applying, the optimal number of clusters (two selected) was determined by the elbow method (Fig.2). As the result we got the first cluster consisting of the 19 regions. The second cluster included only one Dnipropetrovsk region. For evaluation of the cluster effectiveness 2 groups of indicators were used: the Davies-Bouldin index (the result is 3.4% for our clustering) and the Euclidean distance (the result is 90.5%). We can conclude that indicators indicate the high quality of the Ukrainian region's clustering and that the clustering is effective.

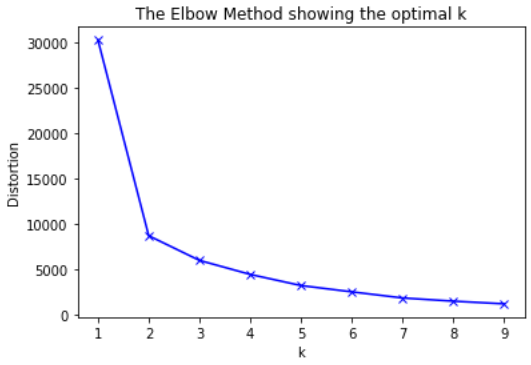

Fig. 2. The Elbow method results

For hierarchical clustering it is important to choose the method for determining distances. The most popular one is Ward's method and it was used. The results for both k-means method and hierarchical clustering can be confirmed, because two clusters are identified in both cases. Analogically to the k-means method, the hierarchical cluster analysis includes only one Dnipropetrovsk region to the first cluster and all other regions to the second one (Fig.3).

The DBSCAN method results depend on the two indicators: the maximum distance between the objects by which they are considered neighbors and the number of objects in the vicinity of the point to be considered as a centroid. When we put the values of these indicators 3 and 2, respectively, we got 2 clusters. But the results are different, than k-means method results, because Zaporizhia region is in one cluster with Dnepropetrovsk region, that is rather logical, because these regions are leaders of industrial development of Ukraine.

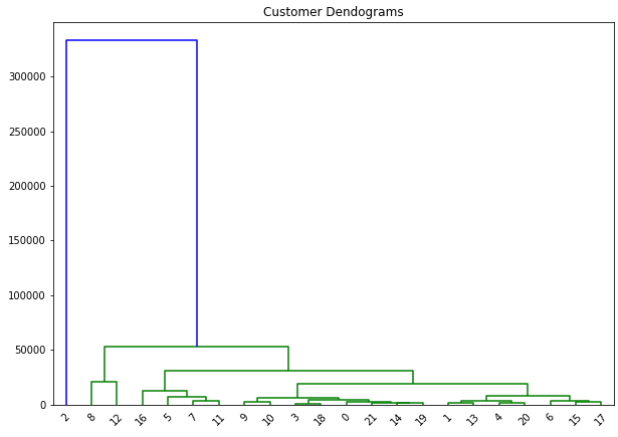

Fig.3. Dendrogram of regions of Ukraine clustering

(1 Vinnytsya, 2 Dnipropetrovsk, 3 Volyn,4 Zhytomyr, 5 Zakarpattya, 6 Zaporizhzhya, 7 Ivano-Frankivs,k 8 Kyiv, 9 Kirovohrad, 10 Lviv, 11Mykolayiv, 12 Odesa, 13 Poltava, 14 Rivne, 15 Sumy, 16 Ternopil, 17 Kharkiv, 18 Kherson, 19 Khmelnytskiy, 20 Cherkasy,21 Chernivtsi,22 Chernihiv)

So the conclusion is that the DBSCAN method shows better results, because of its ability to select industrialized regions

So, we got the same results with k-means and hierarchical clustering, and a little bit different with the DBSCAN method.

4.4. Dimensionality Reduction - Modeling of the firm financial stability with principal component analysis

Classical approaches to the firm's financial stability analysis are based on the calculation of the indicators, whereby the indicators can acquire values from different ranges and have excellent interpretations of the results. Therefore, there is a need for an aggregate indicator to assess the firm's resilience. Another problem is a strong correlation between the indicators, that is, they largely duplicate each other.

The purpose of the investigation is to evaluate the financial stability of a firm with the possibility of comparing such an estimate for different firms using the principal component analysis method.

The information for the analysis of the financial stability of firms is obtained from a private firm "Lviv-Audit", which provides audit services. In order to determine the level of financial stability of firms, the financial statements of five Lviv trade firms were analyzed. The principal component method was used for the evaluation of the firm's financial stability [26].

The following indicators were used to evaluate the financial stability of firms:

1) coefficient of financial independence (autonomy) - shows what proportion of equity in the total balance sheet currency. It takes values from zero to one, normal is considered a value in excess of 0.5. The greater the value of financial autonomy, the more independent the firm from external financing.

2) the coefficient of financial stability - shows what is the ratio between the equity and attracted capital of the firm. If the value of the indicator exceeds one, then its own funds are greater than those attracted, which indicates the financial stability of the firm. The greater the value of the financial stability indicator is greater than one, the more own funds exceed the attracted ones.

3) equity ratio - shows how much working capital falls per unit of stock, the value of 0.6-0.8 considered normative.

4) coefficient of maneuverability of working stocks - shows how much working capital falls per unit of current assets, the regulatory value exceeds 0.1.

5) Equity maneuverability factor - shows how much working capital per unit of equity accounts for.

In the first stage, financial stability indicators from the above list for the five companies were calculated on the basis of the annual financial statements (Balance Sheet and Financial Statement) (Table 3).

Table 3. The financial stability of firms indicators

Firm	x_1	x_2	x_3	x_4	x_5
1	0,162	0,193	0,170	0,137	0,823
2	0,111	0,124	1,958	0,111	1
3	-6,202	-0,861	-27,033	-6,202	1
4	0,000	0,000	0,000	-0,919	-2118,71
5	0,073	0,078	0,124	0,066	0,905

Using the normalized matrix as input, we estimated a correlation matrix (Table 4).

Table 4. Correlation matrix of factors of firms financial stability

Variables	x_1	x_2	x_3	x_4	x_5
x_1	1	0,990	0,998	0,990	0,233
x_2	0,990	1	0,988	0,996	0,120
x_3	0,998	0,988	1	0,989	0,224
x_4	0,990	0,996	0,989	1	0,090
x_5	0,233	0,120	0,224	0,090	1

From the Table 4 follows that there is a correlation between the variables, the maximum value of which is 0.988. Based on the correlation matrix, we can calculate the eigenvectors (Table 5).

Table 5. Eigenvalues of the correlation matrix of factors of firms financial stability

Principal component	Eigenvalues	% total dispersion
Component 1	4,012	80,244
Component 2	0,979	19,584
Component 3	0,007	0,138

From the Table 5 flows that the first principal component (factor 1) explains 80.25% of the total variation, so by including only the first component, we describe with one variable 80.25% of the change in the five variables. This is quite sufficient for practical application. We define the eigenvectors of the correlation matrix (Table 6), which determine the relationship between the variables and the principal components (factors).

Table 6. Eigenvalues of the correlation matrix of factors of firm s financial stability

Variables	Factor 1	Factor 2	Factor 3
x_1	0,499	-0,012	-0,094
x_2	0,495	0,104	0,799
x_3	0,498	-0,003	-0,571
x_4	0,495	0,134	-0,145

x_5	-0,112	0,985	-0,068

Since the components are orthogonal, the removal of the last two factors does not change the eigenvectors of the first factor. As an indicator of financial stability, we use the first major component and get the equation with the usual variables:

$$FS'_1=0,198X_1+1,273X_2+0,045X_3+0,202X_4-0,0001X_5-0,793. \quad (5)$$

This value is an indicator of financial stability. Moreover, a higher value means a better level of the financial stability of the firm. Using the same formula, you can calculate the financial stability for each of the firms (Table 7).

Table 7. Indicator of the firms financial stability

Firm	FS
1	1,1056
2	1,0838
3	-4,0025
4	0,8871
5	0,9260

As can be seen from the Table 8, the financial stability of the investigated firm s varied from 1.1056 to -4.0024, where 0 corresponds to the average value in the line of business, and the change in the greater or lesser side indicates respectively the improvement or deterioration of the financial stability of the firm. According to the data obtained, the highest financial stability indicator of the first firm is 1,106, the second financial stability indicator is 1,084, the fifth - 0,926th, and the fourth 0,887. For these businesses, financial stability indicators are almost at the same level. The worst situation in a third company, its financial stability is -4,003. To increase its financial stability, it is necessary to introduce measures that will help to improve the efficiency of use of fixed assets of the firm; increasing the intensity of use of current assets of the firm; an increase of labor productivity; further increase in sales of goods (products, services); reduction of material operating expenses; expansion of the market for the sale of products (goods, services); attraction of investments (credits), etc.

5 Conclusion

The classic Cobb-Douglas production function (which can be transformed into multiple linear regression) for Ukraine describes rather well the dependencies between the GDP of Ukraine and the number of the working population and capital investment. The scoring model (based on the logistic regression) shows the high predictive power and reliability of the model. The clustering models confirmed a specific role of the Dnipropetrovsk region in Ukraine (according to the competitiveness level of regions of Ukraine). The index of the firm's financial stability for five trade firms was estimated with the principal component analysis method, which allowed to make recommendations for the father firm's development. The objects of investigation represent different levels of economics, what improve the effectiveness of classic machine learning methods in economic research in micro, mezzo, and macro level.

Machine learning has considerable potential for transforming the economy in the near future. Economics researchers need to be sufficiently aware of machine learning methods to be able to use them effectively. It is advisable for economists to analyze the data available, the limitations on their production and analysis, and to increase the interdisciplinary literacy of machine learning methods for successful research

References:

[1] https://www.investopedia.com/terms/m/machine-learning.asp, last accessed 2020/04/01.

[2] Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019) Machine learning in energy economics and finance: A review. Energy Economics, 81, 709-727.

[3] Nehrey, M., & Hnot, T. (2019). Data Science Tools Application for Business Processes Modelling in Aviation. In Shmelova, T., Sikirda, Y., Rizun, N., & Kucherov, D. (Ed.), Cases on Modern Computer Systems in Aviation (pp. 176-190). IGI Global. http://doi:10.4018/978-1-5225-7588-7.ch006.

[4] Newell, R. G., Prest, B. C., Sexton, S. E.: The GDP-temperature relationship: implications for climate change damages. Resour. Future Work. Pap. (2018).

[5] Volkova, N. P., Rizun, N. O., & Nehrey, M. V. (2019) Data science: opportunities to transform education. In Proceedings of the 6th Workshop on Cloud Technologies in Education (CTE 2018), No. 2433, pp. 48-73. CEUR Workshop Proceedings.

[6] Mullainathan, S., Spiess, J. (2017) Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106.

[7] Exterkate, P., Groenen, P. J., Heij, C., & van Dijk, D. (2016) Nonlinear forecasting with many predictors using kernel ridge regression. International Journal of Forecasting, 32(3), 736-753.

[8] Vdovyn, M., Zomchak, L. (2017) Statistical estimation and analysis of foreign trade in EU

and Ukraine. Socio-economic potential of cross-border cooperation. Ivan Franko National University of Lviv, University of Rzeszow, 137-143.

[9] Kim, H. H., & Swanson, N. R. (2014) Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. Journal of Econometrics, 178, 352-367.

[10] Pereira, J. M., Basto, M., & da Silva, A. F. (2016). The logistic lasso and ridge regression in predicting corporate failure. Procedia Economics and Finance, 39, 634-641

[11] Weng, B., Martinez, W., Tsai, Y. T., Li, C., Lu, L., Barth, J. R., & Megahed, F. M. (2018). Macroeconomic indicators alone can predict the monthly closing price of major US indices: Insights from artificial intelligence, time-series analysis and hybrid models. Applied Soft Computing, 71, 685-697

[12] Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: a recent review. Artificial Intelligence Review, 45(1), 1-23.

[13] de Oña, J., de Oña, R., & López, G. (2016) Transit service quality analysis using cluster analysis and decision trees: a step forward to personalized marketing in public transportation. Transportation, 43(5), 725-747.

[14] Harris, T. (2015) Credit scoring using the clustered support vector machine. Expert Systems with Applications, 42(2), 741-750.

[15] Zomchak, L., (2007). Before forecasting analysis of the PFTS index with wavelet technology methods. Problems and prospects of development of the banking system of Ukraine, 2007, 296-304. (in Ukrainian)

[16] Mahajan Shubhrata, D., Deshmukh Kaveri, V., Thite Pranit, R., Samel Bhavana, Y., & Chate, P. J. (2016) Stock market prediction and analysis using Naïve Bayes. Int. J. Recent Innov. Trends Comput. Commun.(IJRITCC), 4(11), 121-124.

[17] Musmeci, N., Aste, T., & Di Matteo, T. (2015) Relation between financial market structure and the real economy: comparison between clustering methods. PloS one, 10(3).

[18] Augustyński, I., & Laskoś-Grabowski, P. (2018) Clustering macroeconomic time series. Econometrics, 22(2), 74-88.

[19] Zomchak, L., Drobotii, Yu. (2020) Regional competitiveness: clustering regions of Ukraine. Modern technologies in the development of economy and human well-being: monograph. Publishing House of University of Technology, Katowice, 20-27.

[20] Siahaan, A. P. U., Mesran, M., Lubis, A. H., & Ikhwan, A. (2017) Association Rules Analysis on FP-Growth Method in Predicting Sales.

[21] French, J. (2017) Macroeconomic forces and arbitrage pricing theory. Journal of Comparative Asian Development, 16(1), 1-20.

[22] Pradhan, R. P., Arvin, M. B., Bahmani, S., Hall, J. H., Norman, N. R. (2017) Finance and growth: Evidence from the ARF countries. The Quarterly Review of Economics and Finance, 66, 136 – 148.

[23] Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2019) Machine learning methods for systemic risk analysis in financial sectors. Technological and Economic Development of Economy, 25(5), 716-742.

[24] State Statistics Service of Ukraine https://ukrstat.org/en

[25] Derbentsev, V., Babenko, V., Pursky, O., Datsenko N., and Pushko, O. (2020). Forecasting Cryptocurrency Prices Using Ensembles-Based Machine Learning Approach. International Scientific-Practical Conference Problems of Infocommunications. Science and Technology PIC S&T'2020

[26] Guryanova, L., Yatsenko, R., Dubrovina, N., Babenko, V. (2020). Machine Learning Methods and Models, Predictive Analytics and Applications. Machine Learning Methods and Models, Predictive Analytics and Applications 2020: Proceedings of the Workshop on the XII International Scientific Practical Conference Modern problems of social and economic systems modelling (MPSESM-W 2020), Kharkiv, Ukraine, June 25, 2020, Vol-2649, 1-5. http://ceur-ws.org/Vol-2649/

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Vitalina Babenko carried out the conzeptualization.
Andriy Panchyshyn has organized formal analysis.
Larysa Zomchak and Maryna Nehrey carried out the literature review, the simulations and the optimizations.
Zoriana Artym-Drohomyretska was responsible for data.
Taras Lahotskyi was responsible for Software.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)