Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2022. Т. 22, вып. 1. С. 130–137
Izvestiya of Saratov University. Mathematics. Mechanics. Informatics, 2022, vol. 22, iss. 1, pp. 130–137
https://mmi.sgu.ru https://doi.org/10.18500/1816-9791-2022-22-1-130-137

Article

Software implementation of ensemble models for the analysis of regional socio-economic development indicators

G. Yu. Chernyshova✉, **N. D. Rasskazkin**

Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia

Galina Yu. Chernyshova, cherny111@mail.ru, https://orcid.org/0000-0002-6464-0408
Nikita D. Rasskazkin, rasskazkin64@gmail.com, https://orcid.org/0000-0002-6796-3616

Abstract. To predict indicators, modern approaches based on machine learning are increasingly being used, as a result, additional tools appear for quantitatively assessing the level of development of socio-economic systems. One of the relevant approaches in machine learning is the use of ensemble methods. The purpose of this study is to develop an approach for processing panel data using special regression models, in particular, the ensembles. An application is presented to implement and compare various regression models, including GPBoost, for panel data used in regional statistics. The application was tested on the example of assessing the innovative potential of Russian regions.

Keywords: panel data, machine learning, boosting, decision tree, regional development
Acknowledgements: This work was supported by the Russian Foundation for Basic Research (project No. 20-010-00465).
For citation: Chernyshova G. Yu., Rasskazkin N. D. Software implementation of ensemble models for the analysis of regional socio-economic development indicators. *Izvestiya of Saratov University. Mathematics. Mechanics. Informatics*, 2022, vol. 22, iss. 1, pp. 130–137. https://doi.org/10.18500/1816-9791-2022-22-1-130-137

Научная статья
УДК 519.688

Разработка приложения для реализации ансамблевых моделей в задаче анализа социально-экономических показателей

Г. Ю. Чернышова✉, **Н. Д. Рассказкин**

Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Россия, 410012, г. Саратов, ул. Астраханская, д. 83

Чернышова Галина Юрьевна, кандидат экономических наук, доцент кафедры дискретной математики и информационных технологий, cherny111@mail.ru, https://orcid.org/0000-0002-6464-0408
Рассказкин Никита Дмитриевич, бакалавр, rasskazkin64@gmail.com, https://orcid.org/0000-0002-6796-3616

Аннотация. При решении задачи анализа социально-экономических показателей применяются актуальные методы машинного обучения, на основе которых создаются современные

инструменты для оценки функционирования социально-экономических систем. Одним из перспективных подходов машинного обучения является использование ансамблевых методов. Целью данного исследования стала разработка подхода для обработки панельных данных с помощью специальных регрессионных моделей, в том числе с применением ансамблей. Представлено разработанное клиент-серверное приложение, позволяющее реализовать и сравнить различные регрессионные модели, в частности модель GPBoost, для панельных данных, используемых в региональной статистике. Приложение апробировано на примере оценки инновационного потенциала российских регионов.

Ключевые слова: панельные данные, машинное обучение, бустинг, дерево решений, региональное развитие

Благодарности: Исследование выполнено при финансовой поддержке РФФИ (проект № 20-010-00465).

Для цитирования: *Chernyshova G. Yu., Rasskazkin N. D.* Software implementation of ensemble models for the analysis of regional socio-economic development indicators [*Чернышова Г. Ю., Рассказкин Н. Д.* Разработка приложения для реализации ансамблевых моделей в задаче анализа социально-экономических показателей] // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2022. Т. 22, вып. 1. С. 130–137. https://doi.org/10.18500/1816-9791-2022-22-1-130-137

Introduction

Panel data are generated in a series of measurements over several periods for the same objects and are often found in the analysis of socio-economic indicators [1,2], in particular, for the presentation of regional statistics. Currently, panel models are being developed, complementing classical methods with modern machine learning approaches [3, 4]. The goals of this study are to design the architecture and develop an application that implements the forecasting of indicators of socio-economic development by R; selection, adjustment, and application of ensemble models. The competitiveness of the region can be interpreted as the efficiency of using the existing regional potential. One of the aspects of the region's competitiveness is the level of innovative development, for the assessment of which the developed application was applied. Various models were used to assess the innovative potential of the regions [5, 6]. To predict the level of regional innovative development using panel regression in the test study, separate financial indicators are used that characterize the formation and use of resources of territories.

1. Applying regression models for panel data

In mathematical representation, panel data is represented as $y_{it}, x_{it}, i = \overline{(1, N)}$, $t = \overline{(1, T)}$, where N is the number of objects, T is the number of observation points. Balanced panel data will be used when describing balanced data models, which means that each time series included in the model has the same number of observation points. Then the total number of observation points is $N * T$, with $N > 1$ and $T > 1$. If $N = 1$, then the data will take the form of a time series, if $T = 1$, then the data will take the form of a slice.

The i-th object can be represented as follows:

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iT} \end{bmatrix} = \begin{bmatrix} X_{i1}^1 & X_{i1}^2 & \cdots & X_{i1}^k \\ X_{i2}^1 & X_{i2}^2 & \cdots & X_{i2}^k \\ \vdots & \vdots & \ddots & \vdots \\ X_{iT}^1 & X_{iT}^2 & \cdots & X_{iT}^k \end{bmatrix} \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{bmatrix}, \tag{1}$$

where y_i is a matrix of dimension $N * T$ containing the predicted parameters, ϵ_i is the error matrix of dimension $N * T$, X is the matrix of the original dataset of dimension, $N * T * k$, k is the number of parameters in each observation, ϵ_{it} is an error term.

Among the panel data analysis models, four main types can be distinguished [7]: a model based on the least-squares method (pooling); a model with a fixed effect (between); a model with a random effect (random); a model with a mixed effect.

The model based on the least-squares method is applicable in the case when there is no heterogeneity in the time series of objects, that is, there are no subsets in the group that are significantly influenced by factors unaccounted for in the sample.

$$y_{it} = X_{it}b + a + \epsilon_{it}, \tag{2}$$

where a is a free term, b is a regression coefficient.

In a fixed-effect model, a free a_i member can take different values for each object of the group. In this case, the regression equation has the form:

$$y_{it} = X_{it}b + a_i + \epsilon_{it}. \tag{3}$$

In a model with random effects, a_i are random variables. In such model, a_i are no longer interpreted as the values of some fixed parameters and are not subject to evaluation. In this case, the model has the form:

$$y_{it} = X_{it}(b + u_k) + a_i + \epsilon_{it}, \tag{4}$$

where u_k represents the random effects.

The random effects model is used in cases where the objects of the study are considered as a subset of the total sample. The advantage of random effects is that it is possible to include time-invariant variables.

The concept of basic models is introduced in the ensemble model of machine learning. Basic models are used to form more complex models by combining several of them. As a rule, the basic models have either a high offset or a large spread. In such cases, the idea of ensemble methods is to reduce the bias and/or spread of such basic models by combining them into one ensemble model that achieves better results. The use of tree boosting as a method of forming ensemble models allowed us to build more accurate models for various predictive tasks [8].

It is proposed to use different approaches for models based on panel regressions [9–11]. In this study, it is proposed to use the GPBoost algorithm based on gradient tree-boosting and mixed effects model [12].

For a mixed model, it is assumed that the predicted variable y is the sum of a potentially nonlinear function $F(X)$ and random effects:

$$y = F(X) + Zb + \epsilon, \quad b \backsim \mathbb{N}(0, \Sigma), \quad \epsilon \backsim \mathbb{N}(0, \sigma^2 I_n), \tag{5}$$

where y is response variable, function F presents fixed effects; X, Z are fixed and random effects predictor variable matrices, ϵ is the independent error term.

In the GPBoost algorithm, the functions $F(X)$ are constructed using an ensemble of trees. This boosting algorithm aims to minimize the risk functional $R(\widehat{F}(X), \Theta)$:

$$R(\widehat{F}(X), \Theta) : (\widehat{F}(X), \Theta) \longrightarrow L(y > t : \Theta)_{F=F(x)}, \tag{6}$$

where $\widehat{F}(X)$ is a linear span of a set of base learners, in our case decision trees, where $L(y > t : \Theta)$ is the negative log-likelihood of the model.

SMAPE and RMSE metrics will be used to evaluate the models. The choice of the SMAPE metric is due to its simple interpretation, as well as its resistance to outliers and the magnitude of the actual value. RMSE is a classical method of estimating the accuracy of regression, by which it is possible to identify the SMAPE bias in case of its occurrence.

2. Application development for panel regression modeling

To develop an application for analyzing socio-economic indicators using ensemble models, a client-server architecture has been chosen. When implementing the algorithms, the R 3.6 platform was used due to the presence of a wide range of specialized libraries for data analysis. The narrow specialization of R affects the limited functionality of the language in other areas, in particular, the server frameworks developed for R are inferior to their counterparts in other platforms. For this reason, the server-side is implemented using Python version 3.9.

In the process of designing an application to implement predictive models, it was proposed to use the Flask framework as a server, the PyQt5 framework for implementing the user interface, and Docker for deploying the server. This approach will ensure that the front-end and R packages are interoperable.

The developed software has an application with a user interface that sends HTTP requests to the REST API server and visualizes the received responses. The submitted requests are received by the Docker container and forwarded to the virtual port of the Ubuntu operating system. Then the request is processed by the Gunicorn server and passed to the Python application using the Flask framework. The data is converted by rpy2 library to the R data format. The converted data is sent to the client application and visualized.

The functionality of the developed application involves the following stages of data analysis: data input in CSV format; implementation of requests for data sampling by periods and objects; application of statistical tests to assess the heteroscedasticity of individual and temporal effects; selection of a predictive algorithm; setting the parameters of tree boosting.

The GPBoost model has several hyperparameters, which allows you to configure the model to increase its accuracy. The most significant parameters are the number of decision trees of the ensemble; the learning step; the maximum depth of the tree; and the minimum number of objects in the leaf of the decision tree. RMSE and SMAPE metrics will be used to evaluate panel data regression. The choice of the SMAPE metric is due to its simple interpretation, as well as its resistance to outliers and the magnitude of the actual value. RMSE is a classic method for estimating the accuracy of regression, which can reveal the SMAPE bias in case of its occurrence.

The interface of the developed application has three tabs: Data, Analysis, Regression. The Data tab allows you to select a file with panel data in CSV format and select the columns responsible for the object ID and period. The Analysis tab allows the user to visualize data panels by time periods or by objects (region identifiers are specified as objects), check data heteroscedasticity, and conduct a series of tests to select a panel data model (Figure). The Breusch-pagan, Honda, King and Wu, F test, Hausman test are used in the application [13]. The Regression tab allows the calculation of forecast values by the implemented algorithms including GPBoost.

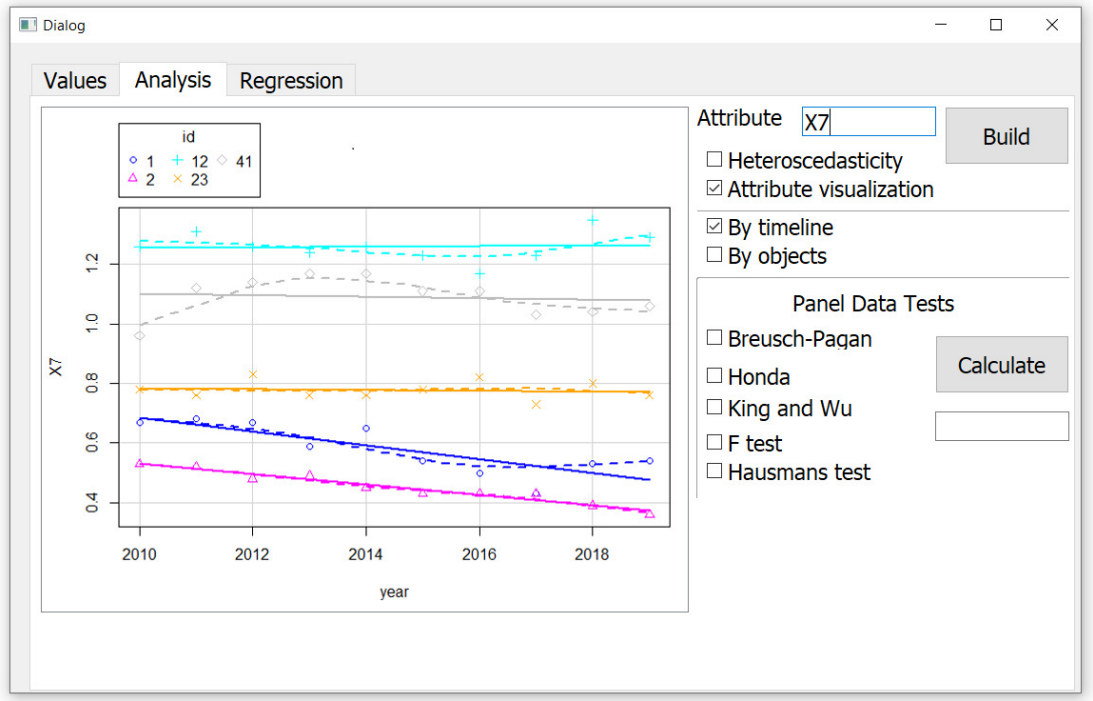

Figure. Application interface for implementing data analysis stages (color online)

The graphs shown in Figure are used for preliminary estimation of the data variability relative to different objects and different periods in advance. The architecture used makes it possible to implement the addition of new forecasting methods, use the application on different user platforms, and integrate the developed modules into existing information and analytical system.

3. Application of the developed application for regression analysis of the innovation potential of regions

To experiment, a sample has been used from the information analysis system FIRA [14]. This resource specializes in the aggregation and provision of information about various economic entities of the Russian Federation. The resulting sample contains information about the socio-economic indicators of 81 Russian regions. Each region is described by 28 indicators. The data are given for the period 2010–2019. The following indicators were selected for the training sample: the volume of innovative goods in the region, the ratio of debt and equity, the availability of own working capital, current liquidity, the share of loans and borrowings in short-term liabilities, capital return, asset

turnover. The volume of innovative goods in the region was chosen as the predicted variable because this indicator reflects the level of innovative regional development. The resulting sample does not contain missing values. This fact suggests that the sample is balanced. Each region as a sampling object has only 10 dimensions per the data for 2010–2019. The same set of parameters is used for all regions included in the sample, which makes the sample homogeneous. The property of balance and homogeneity allows using the resulting sample with different panel models.

The following way of the computational experiment is proposed. The GPBoost ensemble algorithm has been applied with different parameters of base learners. Boosting with trees as base learners has several tuning parameters. Arguably the most important one is the number of boosting iterations (number of trees). Other tuning parameters include the learning rate, the maximal tree depth, the minimal number of samples per leaf, the number of leaves.

Applied models differed by the values of the standard parameters of the algorithm, namely: p_1 is the number of decision trees of the ensemble; p_2 is the maximum depth of the tree; p_3 is the minimum number of objects in the leaf of the decision tree. The constructed models were compared by the RMSE and SMAPE metrics, as well as by the time spent on training the model. To experiment, the sample is divided into two parts: the period 2010–2017 for training models, 2018–2019 for the accuracy assessment of the constructed models. The computational experiment is divided into 2 stages. At the first stage, the number of trees (p_1) and the maximum depth of the tree (p_2) were simultaneously selected, at the second, the minimum number of objects in the tree leaf (p_3). Such selection of parameters p_1 and p_2 is because the number of decision trees and the maximum depth of the tree are largely interrelated parameters. For this reason, a model constructed using independently determined best values of parameters p_1 and p_2 will not necessarily show the best result. At the first stage, 6 models were built using different parameters p_1 and p_2 (Table 1). According to the results of the first part of the experiment, it turned out that the $M_{1.2}$ model showed high accuracy for both measures. However, it is worth noting that the $M_{1.5}$ model is slightly inferior in accuracy to $M_{1.2}$. For this reason, for further investigation of the GPBoost parameters, the values of the parameters p_1 and p_2 used in the $M_{1.5}$ model will be used, namely, 7 and 5, respectively.

Table 1

Selection of parameters p_1 and p_2

Model	p_1	p_2	RMSE	SMAPE
$M_{1.1}$	5	3	1.153	0.882
$M_{1.2}$	7	3	0.650	0.447
$M_{1.3}$	9	3	0.679	0.456
$M_{1.4}$	5	5	1.135	0.866
$M_{1.5}$	7	5	0.659	0.451
$M_{1.6}$	9	5	0.722	0.478

At the second stage of the experiment, models $M_{2.1}$, $M_{2.2}$, and $M_{2.3}$ were constructed. These models have identical parameters with model $M_{1.5}$, p_1 and p_2, but have different values of parameter p_3 (Table 2).

Table 2

Selection of parameter p_3

Model	p_3	RMSE	SMAPE
$M_{2.1}$	5	0.691	0.469
$M_{2.2}$	10	0.659	0.451
$M_{2.3}$	20	0.674	0.458

Changing different values of parameter p_3 did not give a significant accuracy increase. For the further experiment, model $M_{2.3}$ was chosen, which showed high accuracy of forecasting by RMSE and SMAPE. Parameters p_1, p_2, p_3 have values 7, 5, and 20 respectively.

The obtained GPBoost model was compared with classical panel regression methods (Table 3). GPBoost model shows the highest prediction accuracy. The hybridization of methods, in particular, the using a set of decision trees into a single ensemble model gave more accurate models.

Table 3

Evaluation of panel regression models

Model	GPBoost	Pooling	Between	Random
RMSE	0.674	1.327	1.170	0.721
SMAPE	0.458	0.930	0.901	0.512

This model can be used to predict both the volume of innovative products and other indicators of regional development. The proposed approach based on hybrid methods of panel regression and decision trees provided a sufficiently high accuracy of the forecast relative to classical panel methods.

Conclusion

A developed client-server application providing implementation and comparison of various forecasting algorithms for panel data has been presented. The advantage of panel methods is the ability to account for bias due to unselected factors or undersampling. In addition, panel models have superior predictive accuracy with a relatively small sample size, which is typical for forecasting tasks related to regional statistics.

This application allows you to use panel data models with fixed and random effects, as well as a hybrid ensemble GPBoost model. The GPBoost model allows you to combine decision tree boosting methods and panel regression. The architecture of the application assumes the use of different R packages for data analysis. The application server-side is implemented on the Python platform. The user interface is realized by the PyQt5 framework. The extended functionality of the application provides loading and selection of data, execution of statistical tests for heteroscedasticity, panel model fitting, using the selected model to obtain a predicted value. A feature of this application is the possibility of preliminary analysis of a sample with panel data, comparison of various models by some widespread error measures.

The represented application is focused on the implementation of regression analysis of various aspects of regional development based on data with a panel structure to obtain a comprehensive assessment and forecast of the region's competitiveness as a whole.

References

1. Aivazian S. A. *Metody ekonometriki* [Methods of Econometrics]. Moscow, INFRA-M, 2019. 512 p. (in Russian).
2. Greene W. H. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ, Prentice Hall, 2003. 1026 p.
3. Hajjem A., Bellavance F., Larocque D. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 2014, vol. 84, iss. 6, pp. 1313–1328. https://doi.org/10.1080/00949655.2012.741599
4. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Liu T. Y. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing system*, 2017, vol. 30, pp. 3146–3154.

5. Firsova A., Chernyshova G. Efficiency analysis of regional innovation development based on DEA Malmquist index. *Information*, 2020, vol. 11, no. 6, 294. https://doi.org/10.3390/info11060294
6. Veshneva I., Chernyshova G. The scenario modeling of regional competitiveness risks based on the Chapman-Kolmogorov equations. *Journal of Physics: Conference Series (JPCS)*, 2021, vol. 1784, iss. 1, 012008. https://doi.org/10.1088/1742-6596/1784/1/012008
7. Gurka M. J., Kelley G. A., Edwards L. J. Fixed and random effects models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2011, vol. 4, iss. 2, 181–190. https://doi.org/10.1002/wics.201
8. Breiman L., Friedman J. H., Stone C. J., Olshen R. A. *Classication and Regression Trees*. 1st ed. New York, CRC Press, 1984. 368 p. https://doi.org/10.1201/9781315139470
9. Laird N. M., Ware J. H. Random-effects models for longitudinal data. *Biometrics*, 1982, no. 38, pp. 963–974.
10. Pinheiro J., Bates D. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006. 528 p.
11. Rasmussen C. E., Williams C. K. J. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. 266 p.
12. Sigrist F. Gaussian Process Boosting. *arXiv preprint arXiv*, 2020.
13. Baltagi B. H. *Econometric Analysis of Panel Data*. 6th ed. Chichester, John Wiley & Sons, 2021. 436 p.
14. *Information Analysis System FIRA PRO*. Available at: https://pro.fira.ru (accessed 15 September 2021).

Поступила в редакцию / Received 24.11.2021
Принята к публикации / Accepted 21.12.2021
Опубликована / Published 31.03.2022