EVALUATING MACHINE LEARNING MODELS FOR HEALTHCARE SERVICES EFFICIENCY

AVALIANDO MODELOS DE APRENDIZADO DE MÁQUINA PARA EFICIÊNCIA DE SERVIÇOS DE SAÚDE

Ani Sukiasyan
Candidate of Economic Sciences
Department of Mathematical Methods in Economics
Plekhanov Russian University of Economics
Moscow, Russia
sukiasyan.ag@rea.ru

Resumo. O artigo é dedicado ao problema de avaliação da qualidade dos cuidados de saúde utilizando algoritmos de aprendizagem de máquina. Propõe-se avaliar o termo categórico "qualidade dos cuidados de saúde" como variável discreta com base em dados sobre o resultado do tratamento, fornecidos por um dos principais hospitais da Rússia. É fornecida uma análise preliminar das variáveis dependentes e explicativas. São analisadas as estatísticas dos resultados do tratamento em função da idade e sexo dos pacientes. A necessidade de redução do conjunto de dados inicial é justificada. São propostas as formas de correção do desequilíbrio no conjunto de dados, como SMOTE e Tomek Links. Com base nos dados corrigidos, foram projetados modelos logit ordenados e Random Forest. É apresentada a análise comparativa de vários modelos. As desvantagens de cada modelo são explicadas. O modelo logit ordenado em dados balanceados permitiu determinar os fatores que têm maior impacto positivo no resultado do tratamento, bem como o modelo Random Forest tornou as previsões tão precisas quanto possível. Com base nos resultados obtidos, são feitas recomendações para melhorar a eficiência dos cuidados de saúde no âmbito de um determinado hospital.
Palavras-chave: Eficiência em Saúde, Aprendizado de Máquina, Modelo Logit, Random Forest, SMOTE, Tomek Links.

Abstract. The article is dedicated to the problem of assessing the quality of healthcare using machine learning algorithms. It is proposed to evaluate categorical term "healthcare quality" as discrete variable on the basis of data on treatment result, provided by one of the major hospitals in Russia. Preliminary analysis of both dependent and explanatory variables is provided. The statistics of results of treatment depending on the age and gender of patients are analyzed. The necessity of reducing the initial dataset is justified. The ways of correction the imbalance in dataset is proposed, such as SMOTE and Tomek Links On the basis of corrected data ordered logit and Random Forest models were designed. The comparative analysis of various models is presented. The drawbacks of each model are explained. The ordered logit model on balanced data allowed to determine factors, that have greatest positive impact on the treatment result as well as Random Forest model made the predictions as accurate as it is possible. On the basis of results obtained, the recommendations to improve the healthcare efficiency in the framework of particular hospital are made.

Keywords: Healthcare Efficiency, Machine Learning, Logit Model, Random Forest, SMOTE, Tomek Links.

INTRODUCTION

Healthcare system is one of the most complicated spheres for the state management due to its peculiarities. However, it is crucial to develop the most of possible efficient control and regulation system for healthcare, as imbalanced regulation will cause unjustified expenses in this sphere in short term time period. Moreover, in long term time period the absence of optimal course of action will cause also deterioration of the quality of healthcare for patients, which, as a consequence, will follow to decrease in active and healthy years of living or premature deaths among population of different ages, especially among working-age population, which will follow decline in all spheres of countries economy (Tikhomirov & Tikhomirova, 2021; Tikhomirova & Sukiasyan, 2021; Shariati et al., 2013; Voronkova et al., 2022). For Russia this problem is nowadays one of the most relevant, since the country is solving problems of misuse of budgetary funds, including those dedicated to healthcare sector. In addition, due to the demographic crisis (Tikhomirov et al., 2019), Russia, along with measures to stimulate the birth rate, is solving the tasks of reducing the level of premature mortality of the population. The reduction of mortality due to various diseases affects the quality of medical care, which must be maintained at the highest possible level in conditions of a limited budget. It is possible to assess the quality of medical care by analyzing the efficiency of patient treatment depending on various factors. To solve this problem, it is necessary to build a mathematical model that will include variables (Farhud & Mojahed, 2022), that have the greatest impact on the health of patients. This will allow to improve the effectiveness of treatment by adjusting the values of certain variables. In other words, to monitor and make decisions to influence these factors, which in turn will lead to improve the

patient's health status. As a result, this will have a positive impact on the effectiveness of the healthcare system in the country as a whole.

The continuous development of intelligent systems aims to provide better and more efficient reasoning using the collected data. This use is not limited to retrospective interpretation, i.e., the diagnostic presentation of conclusions can also be extended to probabilistic interpretation that provides early prognosis. That said, the clinicians who can be assisted by these systems themselves stand in the inter-clinical gap subject to deep technical scrutiny. What they lack is a clear starting point from which to approach the world of machine learning in medicine.

Collected medical data can be analyzed using different methods at different levels. The first level of patient acquisition is the data that conventional alarm systems can help collect regarding when values are outside the normal range, as is the case with electrocardiographic devices (heart tape). In the second level, different data sources are collected, combined and processed so that the result can be used. Input to another type of system that provides differential propositional recognition and conclusions based on a set of rules. By navigating a tree-like hierarchy using the given data, these systems can help arrive at a logical explanation of the input signals. These rule-based systems are called "expert systems".

Systems experts learn from experience to imitate the abilities of human experts by making decisions. These systems often have the ability to answer questions that begin with "what", "how", "where". At the same time explain the reasoning of their decisions. Another basic feature of these systems is that it integrates new experiences, thus enriching and increasing their knowledge. This, in turn, improves decision-making abilities. The prototype of these systems is the MYCIN system. These systems are based on a data transformation process that provides diagnosis and conclusions that are firmly established in the "Data-Information-Knowledge-Intelligence" model. Current goals are the continuous development of intelligent systems to provide better reasoning and efficient use of collected data. The goal is to enable decision-making systems to provide prospective implementation and early prognostication (as opposed to a retrospective approach, where systems only provide diagnosis and conclusions). These technical achievements are well documented, however, providing a simple tutorial for clinicians seeking to understand and further demonstrate the true status of this technology and its potential uses.

Intelligence is one of those terms that resist any definition. Attempting simple web browsing can lead us to hundreds of definitions that vary according to individual perspectives (philosophy, biology, psychology, mathematics, computer science).). However, for the sake of the progress of this paper, we will try to collect many definitions available in the literature. Intelligence is the ability to create adaptive designs, solve problems, or create or create a product that has value in a particular culture or in business. From association, memorization, reasoning, understanding, abstraction, conceptualization, approximation, systematization and logical deduction. These elements are trained to acquire new knowledge from known facts, on the other hand, artificial intelligence refers to the system's ability to correctly interpret external data, learn from this data, and use this learning to achieve Be flexible with specific goals and tasks. Machine learning is when computers are used to apply statistical models to data. It is a subfield of artificial intelligence, where computer programs (algorithms) learn relationships between input and output data. Three categories of machine learning algorithms can be distinguished:

- Supervised learning. In supervised learning, computer programs learn associations through the analysis of data samples defined by a human expert observer in a process called training. Once the associations have been learned, they can be used to predict future instances, for example in a process called testing.
- Unsupervised learning. In unsupervised learning, computer programs learn associations without externally defining the associations in the data. They learn that it is often used for clustering, that is, extracting undiscovered correlations in input data in such a way as to form subsets of data that share common characteristics.
- Reinforcement learning. In reinforcement learning, the system learns how to behave based on reward/punishment. Punishment can be considered as a negative reward signal that reinforces an action to prevent its delivery.

The analysis of data on treatment results

The dataset, used in this research contained data on 1 764 020 services, provided to patients of one of the major hospitals in Russia and their health state before and after treatment and main personal metrics (Donoso et al., 2022). The list of explanatory variables contains the results of treatment in various profiles (cardiological, cardiac surgery, gastroenterological, hematological, gynecological, etc.) as well as main

characteristics of patients as age, gender, cost of services provided, duration of inpatient period, source of funding (compulsory medical insurance (CMI), voluntary medical insurance (VMI), high-tech medical care (HMC), paid medical services (PS), etc.), type of patient (scheduled or emergency) (Ferrer et al., 2022; Cardenas, 2023). To analyze such dataset, in the framework of this research machine learning models were selected, as they are much efficient and less time consuming as classic regression, which are not applicable in case of nonlinear relations, multicollinearity or enlarged number of variables (Hastie et al., 2009; James et al., 2017). Furthermore, machine learning models are the best choice in case if dependent variable is discrete, which will be discussed below.

Before the model was designed, it was necessary to provide preliminary data analysis and manipulations in order to make it more homogeneous since data contained the records of patients, which took treatment in different departments of the hospital, such as therapy, oncology, cardiology. Thus, the result of treatment in different departments depends on not only the quality of medical service, but also specifics and severity of the disease and other social, demographic and economic conditions.

Taking into account the above it was decided to divide the dataset into smaller sets on the basis of departments and analyze each of them separately. This article observes the analysis and modelling of the efficiency of treatment of data, based on the therapy department dataset. Besides, the dataset was cleared from data, which contained outliers, as well as from missing records. The following table (see

Table **1**) gives a summary of all outcomes for therapy dataset.

The final dataset, used in a framework of this research, contains information about the age, gender, type of patient, source of funding, treatment profile, treatment result, total cost of medical services and the number of days spent in the hospital. For the further analysis, patients who spent no more than 30 days in the hospital were considered.

Table 1. Distribution of patients of therapy department by treatment results.

Treatment Result	Quantity	%
Improvement	14005	83.11
No changes	1182	7.01
Recovery	1115	6.62
Not specified	432	2.56
Transferred to another hospital	70	0.42
Fatal outcome	31	0.18
Discharged with a deterioration	15	0.09
Transferred to the inpatient	1	0.01
Transferred to the outpatient	0	0

It seems interesting to discuss specifics of variables mentioned above. Concerning the result of treatment, which is the dependent variable in a model below, preliminary analysis of it, shown in Figure 1 indicates that, first of all, the number of patients with improvement of both genders per 1000 persons significantly exceeds the number of recovered patients or with no changes. Besides, the proportions of females and males with improvement, as well as with no changes are comparably the same, while the number of recovered females is 1.6 times higher, than the number of recovered males, which may indicate that the health state of males before the treatment took place was worse, than for females and corresponding treatment was more efficient for females than for males.

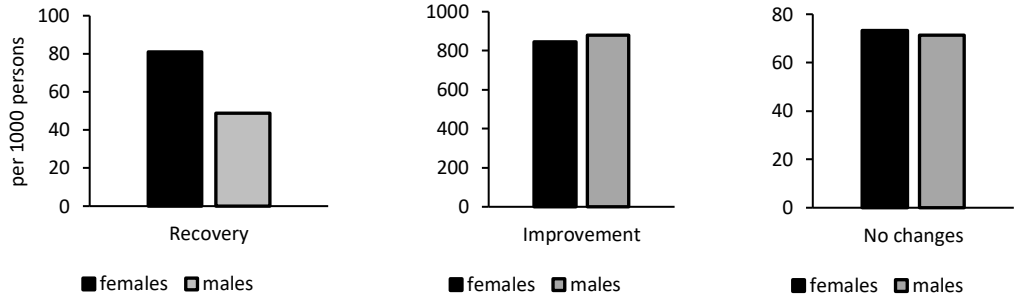

Figure 1. Gender structure of patients with selected outcomes for therapy department data.

THE MULTIVARIATE ANALYSIS OF DATA ON PATIENTS

A more detailed analysis was performed for explanatory variables on patients and services provided for them. In accordance with its results, the following conclusions for patients left the hospital with different treatment results.

Recovered patients

Females. The age distribution is normal with the peak in the age group from 30 to 39 years. Mostly these are scheduled patients – 75%. The majority (52%) were served on the basis of VMI funds. The main treatment profiles were gynecological, otorhinolaryngological, surgical and urological. More than 50% of patients spent less than 5 days in the hospital. The cost of services for more than 40% of patients ranged from 20 to 40 thousand rubles.

Males. The age distribution is normal with the peak in the age group from 30 to 39 years. The proportion of scheduled patients is 69%. The most (2/3) of the patients were admitted via the VMI funds. The main treatment profiles were otorhinolaryngological, surgical and urological. Also, 62% of patients spent 5 to 9 days on treatment. The cost of treatment of the main proportion of patients was within 60 thousand rubles. The average cost is 44 thousand rubles.

Patients left the hospital with improvement

Females. The age distribution is normal with the peak in the age group from 60 to 69 years. These are mainly scheduled patients were served through the funds of CMI (44%) and VMI (23%). The main treatment profiles were orthopedic, rehabilitation, oncological, gynecological, rheumatological, neurological, cardiological, and therapeutic. 46% of the patients spent up to 5 days in the hospital. The cost of medical services of 80% of patients was up to 50 thousand rubles. Average cost of treatment was 39 thousand rubles.

Males. These are mostly people aged 50 to 69 years. Most of the patients were admitted as scheduled and their treatments were paid by the CMI funds (42%). The main treatment profiles were rehabilitation, orthopedic, craniological, neurological, rheumatological, therapeutic, oncological, cardiac surgical, hematological. It is obtained that 75% of the patients were hospitalized for up to 9 days. The cost of treatment of 72% of patients was up to 50 thousand rubles. Average cost of treatment was 49.5 thousand rubles.

Patients left the hospital without changes

Females. Most of the patients belonged to the age group from 40 to 69 years. The majority of patients – 97% – are scheduled. More than half (54%) are patients, whose medical services were paid by CMI fund. The main treatment profiles were oncological (72%), hematological (10%) and neurological (4%). The duration of treatment of 58% of patients did not exceed 4 days. The cost of treatment for 64% of people did not exceed 50 thousand rubles. Average cost of treatment was calculated to be 65 thousand rubles.

Males. A significant proportion of patients (48%) belonged to the age group from 50 to 69 years. The share of scheduled patients was 97%. For this group of patients 36% were served as PS, 25% – by the CMI and 25% – by the VMI. The main treatment profiles were oncological (54%), hematological (18%), neurological (6%), therapeutic (5%) and ophthalmological (3%). The duration of treatment of 56% of people was no more than 4 days on treatment. The cost of treatment of 70% of patients did not exceed 50 thousand rubles. Average cost of treatment for this group was 52 thousand rubles.

MACHINE LEARNING MODELS FOR TREATMENT EFFICIENCY

In accordance with the frequencies obtained and presented in
Table **1** above only records concerning the first three treatment outcomes were representative for modelling. Thus, the corresponding records were selected and following denotations for outcomes were introduced (see
Table **2**):

Table 2. Denotations of discrete depended variable values.

Treatment Result	Denotation
Recovery	2
Improvement	1
Without change	0

In order to clear dataset from outliers the records, which included services with a cost greater than 200 thousand rubles. Moreover, the records with greater frequencies of source of funding and treatment profiles were selected. The discrete variables for source of funding take following values:

$$VMI = \begin{cases} 1, & \text{if the funding source is voluntary medical insurance} \\ 0, & \text{otherwise} \end{cases} \quad (1.1)$$

$$CMI = \begin{cases} 1, & \text{if the funding source is compulsory medical insurance} \\ 0, & \text{otherwise} \end{cases} \quad (1.2)$$

$$PS = \begin{cases} 1, & \text{if the funding source is paid by patient medical service} \\ 0, & \text{otherwise} \end{cases} \quad (1.3)$$

By the treatment profile there were four profiles selected: Rehabilitation, Orthopedic, Oncological and Gynecological. The corresponding dummy variables were obtained similarly to those above (see formulas(1.1),(1.2) and(1.3)).

Thus, the final sample included 15850 rows. The main reason of reducing the number of observations is, as it was mentioned above, is to obtain balanced sample (Jamalpour & Yaghoobi-Derab, 2022). This is required procedure as machine learning models, based on linear classification, are very sensitive to any outlier (He & Garcia, 2009; Kavrin & Subbotin, 2018). If the sample is not balanced and contains outliers or classes with a small number of objects, the model will be less accurate and the forecasts will have little precision (Shunina et al., 2015).

Ordered discrete choice model for evaluating the efficiency of healthcare

The first type of models applied to existing dataset are ordered discrete choice models. In accordance with definition, the dependent variable in this type of models takes values, which are an ordered sequence (Tikhomirova & Sukiasyan, 2021; Tikhomirova & Sukiasyan, 2020):

$$y_i = \begin{cases} 0, & z_i \le b_1 \\ 1, & b_1 < z_i \le b_2 \\ \dots & \\ k, & z_i > b_2 \end{cases} \quad (1.4)$$

In formula (1.4) b_i is the threshold values for classes, k is the number of classes, z_i is the linear combination of explanatory variables X_{il} with particular wages a_l (l is the number of explanatory variables):

$$z_i = \alpha_0 + \sum_{l=1}^{m} a_l X_{il} + \varepsilon_i \quad (1.5)$$

Thus, the ordered logit model is:

$$\begin{cases} p_i = P\{y_i = j \mid X_i\} = \dfrac{1}{1+e^{-z_i}}, & j = \overline{0,k} \\ z_i = \alpha_0 + \sum_{l=1}^{m} a_l X_{il} + \varepsilon_i \end{cases} \quad (1.6)$$

In the formula (1.6) p_i is the probability of object i to be included to class j in accordance with the values of explanatory variables X_{il}.

The confusion matrix of the ordered logit model, given in indicates, that the best predictions are obtained for the class of patients, who left the hospital with improvement (97.6%), while the other two classes ("no changes", "recovery") were characterized with the accuracies 1.13% and 20.0% respectively. This result can be explained simply by the predominance of class "improvement" in dataset: 86% of patients, included in the sample, left the hospital with improvement.

To overcome this problem two modifications were applied (King & Zeng, 2001). The first modification provides for adding wages to classes in accordance with their frequency in the sample:

$$w_i = \frac{n}{k \cdot n_i} \qquad (1.7)$$

In formula (1.7) n is the sample size, k is the number of classes, n_i is the number of objects in class i.

Implementation of this modification to the sample under consideration led to worse accuracy, than it was obtained for the model without weighting the classes – 77% taking into consideration that prevalence of the class 1 in sample was 86%.

The second modification suggests to overcome the problem with imbalanced sample using the SMOTE tool (Synthetic Minority Oversampling Technique) (Chawla et al., 2002). After the oversampling, obtaining the subsample of 40893 observations and implementing the ordered logit model to it, the accuracy obtained was almost the same as for the model with wages – 76,8%. But it should be mentioned, that the quality of this model was significantly better than for the model with weights, as the percentage of correctly predicted outcomes in classes was equally the same. The analysis of quality metrics for models are given in following table (see

Table **3**).

Table 3. Percentage of correctly predicted outcomes of ordered choice models.

Classes	Ordered choice model	Weighted ordered choice model	Ordered choice model after SMOTE
0 (no changes)	1.13	76.94	72.89
1 (improvement)	97.59	83.79	73.61
2 (recovery)	20.00	1.40	83.90

But not only is the accuracy of ordered choice models to be analyzed, but the marginal effects. As the quality of the model after SMOTE is the most acceptable in comparison with the others, the marginal effects were calculated for it. In accordance with the values of marginal effects in the mean point of the sample, variables, which contribute to the treatment profiles Oncological and Gynecological and the funding source VMI, have the greatest effect on the efficiency of treatment.

It seems interesting to consider the marginal effects in more detail. The patient's belonging to the oncological profile increases their probability of leaving the hospital without changes by 59% and reduces the probability of leaving the hospital with improvement by 42% and with recovery by 16.9%. Belonging to a gynecological profile, on the contrary, increases the probability of leaving the hospital with recovery by 51.6% and reduces the probability of leaving the hospital without changes or with improvement (by 15.7% and 35.9%, respectively). Belonging to a rehabilitation or orthopedic profile has a much smaller impact on the probability of treatment results: these factors increase the probability of leave the hospital with recovery (by 2.6% and 5%, respectively) and reduce the probability of discharge with or without improvement.

In turn, if the funding source is VMI it increases the probability of leaving the hospital with recovery by 34% and reduces the probability of leaving the hospital with improvement and without change (by 20.5% and 13.5%, respectively). Funding through CMI has less influence on the probability of a treatment result, but it is worth noting that this factor increases the probability of leaving the hospital with no changes by 1.29% and reduces the probability of leaving the hospital with recovery by 1.25%. Funding through PS channel also increases the probability of leaving the hospital without changes (by 4.8%) and reduces the probability of leaving the hospital with recovery (by 4%).

On the basis of ordered logit model after SMOTE it seems reasonable to clarify the boundaries of classes and thereby improve the quality of classification. For this case Tomek Links method was used, however its implementation to a sample had little improvement on results of modelling (see

Table **4**) (Sherafatizangeneh et al., 2022). This indicates, that the to improve the accuracy of prediction on this model the other types of machine learning models should be implemented.

Table 4. Percentage of correctly predicted outcomes of ordered choice models.

Classes	Ordered choice model after SMOTE	Ordered choice model after SMOTE + Tomek Links
0 (no changes)	72.89	72.87
1 (improvement)	73.61	74.61
2 (recovery)	83.90	84.22
Accuracy of model	76.80	77.30
McFadden R^2	0.43	0.44

After implementing the proposed algorithm to initial dataset the results were obtained, which have the same explanation as above (see

Table **5**). The models demonstrate almost the same results with the proportional loss in accuracy due to drawbacks of initial dataset.

Table 5. Percentage of correctly predicted outcomes of ordered choice models designed for initial dataset.

Classes	Ordered choice model after SMOTE	Ordered choice model after SMOTE + Tomek Links
0 (no changes)	66.58	67.01
1 (improvement)	73.61	72.50
2 (recovery)	80.56	81.12
Accuracy of model	73.57	72.68

Random Forest for assessing the efficiency of healthcare

Since the Random Forest model is less sensitive to outliers and the number of variables, we will train this model on 3 sampling options:

- S1: a sample, that includes patients, who spent less than 30 days in the hospital, whose cost of treatment was up to 200 thousand rubles. The data on the funding source and treatment profile here were reduced to dummy variables in accordance with previous subsection.
- S2: a sample, that includes patients, who spent less than 30 days in the hospital, whose cost of treatment was up to 200 thousand rubles. The data on the funding source and treatment profile here were reduced to dummy variables without exceptions. The resulting sample contains 5 dummy variables for the admission channel and 22 for treatment profiles.
- S3: a sample, that includes patients, who spent less than 30 days in the hospital. The data on the funding source and treatment profile here were reduced to dummy variables without exceptions, similar to the above.

The quality metrics of models was considered to be the accuracy, obtained using stratified cross-validation, which divides the sample into 5 parts. The model on sample S3 has the highest accuracy (see Table 6).

Table 6. Quality metrics of Random Forest models, designed on samples S1, S2, S3.

Sample	Accuracy
S1	87.52
S2	89.19

	S3	89.36

Despite greater values of accuracy of models, the same problem as in case of ordered logit model occurs here. The accuracy of correctly predicted classes, obtained for example on the basis of sample S3, for the small-number classes 0 and 2 are insufficient (40.06% and 62.33%, respectively), while the accuracy of the prediction of class 1 is 95.42%. The reason of such results is also the same – imbalanced dataset.

Implementation of the weighting and SMOTE to solve the mentioned above problem brings to following results (see

Table **7**).

Table 7. Quality metrics of Random Forest models, designed on samples S1, S2, S3 (accuracy).

Sample	Random Forest	Random Forest with weighting	Random Forest after SMOTE
S1	87.52	87.49	91,73
S2	89.19	89.22	94,42
S3	89.36	89.33	94,44

The Random Forest and Random Forest with weighting the samples demonstrate almost the same results. Thus, Random Forest with weighting does not improve the quality of prediction. While implementation of SMOTE tool significantly improves the accuracy of models as well as the accuracy of classes predictions, which is presented in the

Table **8**.

The procedure of clarifying the boundaries of classes using Tomek Links has some positive impact on the quality of prediction. Thus, the Random Forest after SMOTE + Tomek Links was implemented to dataset S3 in order to demonstrate the final confusion matrix of the model. It should be mentioned, that Tomek Links method reduced the sample, which was created using the SMOTE method, by 3972 observations. Thus, the sample, included 38043 patients was used to design this model. The accuracy of Random Forest model using cross-validation on the new sample is 94.44%, i.e., it does not differ from the accuracy on the extended sample.

Table 8. Quality metrics of Random Forest model, designed on sample S3 (accuracy).

Treatment results	Random Forest	Random Forest with weighting	Random Forest after SMOTE
0	40,06	41,88	97,22
1	95,42	95,57	88,24
2	62,33	61,17	97,89

Consider the confusion matrix of the model (see

Table **9**). The proportions of correctly predicted elements of classes 0 and 2 differ slightly from the proportion of correct ones in the model trained on the SMOTE sample. The proportion of correctly predicted objects of class 1 is greater than this indicator for the model trained on the SMOTE sample by 1%. Indeed, the Tomek Links method improves the classification quality of individual classes, but only slightly.

Table 9. Confusion Matrix of Random Forest model after SMOTE + Tomek Links, designed on sample S3.

	Outcomes	p			Accuracy
		0	1	2	
y_i	0	12329	281	52	97.37
	1	792	11130	567	89.12
	2	75	198	12619	97.88

CONCLUSION

To evaluate and predict the efficiency of healthcare services several tasks were set and solved. Firstly, it was proposed to assess categorial term "healthcare quality" as discrete variable, the values of which indicate the result of treatment. Secondly the drawbacks of dataset on patients were determined. The main was its imbalance, i.e. the presence of outliers and classes with low frequencies, on the basis of which a few manipulations were applied, that made the more appropriate for modelling.

Two types of machine learning models were designed on the basis of modified dataset. Despite of the difference in algorithms of these models, it was noticed, that the more imbalanced is the dataset, the more likely the models to predict the results of treatment of classes with less frequency as the result of treatment of prevailing class. It was pointed out that the fact that high accuracy of model is not enough to conclude the high quality of prediction, as the accuracy of model mostly depends on the size of prevailing class, which is always well predicted due to features of algorithms. Thus, additional tools should be implemented to make the quality of prediction in all classes better.

As a result, implementation of SMOTE + Tomek Links allowed to design ordered logit and Random Forest models on balanced sample and determine, that independent variables explain fully the result of treatment. What is more, in accordance with marginal effects obtained, such variables as the treatment profile and funding source have the greatest positive impact on the treatment result. The most and least efficient profiles were indicated, which allows to correct the regulation of the healthcare system. Taking into account, that the greatest impact on positive treatment result among different ways of funding has VMI it should be decisions made to include greater range of healthcare services to other insurance programs to make them available to extended segments of the patients.

The models can also be used to make an accurate prediction of the result of treatment of new patients in accordance with their type (scheduled or emergency), age, gender, cost of services provided, duration of inpatient period, source of funding.

ACKNOWLEDGMENTS

This research was performed in the framework of the state task in the field of scientific activity of the Ministry of Science and Higher Education of the Russian Federation, project "Models, methods, and algorithms of artificial intelligence in the problems of economics for the analysis and style transfer of multidimensional datasets, time series forecasting, and recommendation systems design", grant no. FSSW-2023-0004.

REFERENCES

Cardenas, C. A. (2023). Sarcomas:"A Comprehensive Review of Classification, Diagnosis, Treatment, and Psychosocial Aspects. Clin Oncol Case Rep 6, 6, 2-6.

Chawla N.V., Bowyer K.W., Hall L.O., W.P. (2002). Kegelmeyer SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Donoso, P. C., Pérez, M. P. S., Aguirre, C. C., Barbosa, A. O., Gómez, C. M. G., Jimenez, A. M., & Nodar, S. R. (2022). Angiosarcoma suprarrenal primario. Reporte de caso. Archivos de Patologia, 3(3), 96-103.

Farhud, D., & Mojahed, N. (2022). SARS-COV-2 Notable Mutations and Variants: A Review Article. Iranian Journal of Public Health, 51(7), 1494.

Ferrer, N. R., Romero, M. B., Ochenduszko, S., Perpiñá, L. G., Malagón, S. P., Arbat, J. R., & Nodar, S. R. (2022). Solitary fibrous tumor of the thyroid. Report of a case with unusual clinical and morphological findings Archivos de Patologia, 3(3), 104-109.

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer, New York

He H., Garcia E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21, 1263–1284. DOI: 10.1109/TKDE.2008.239.

Jamalpour, H., & Yaghoobi-Derab, J. (2022). A review of the philosophy of aesthetics and art based on theoretical and methodological considerations. Revista de Investigaciones Universidad del Quindío, 34(S2), 426-435.

James G., Witten D., Hastie T, Tibshirani R. (2017). An Introduction to Statistical Learning. Springer, New York.

Kavrin D.A., Subbotin S.A. (2018). Methods for quantitatively solving the problem of class imbalance. Radio Electronics, Computer Science, Control, 1(44), 83-90.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137 - 163

Shariati, A., Azaribeni, A., Hajighahramanzadeh, P., & Loghmani, Z. (2013). Liquid–liquid equilibria of systems containingsunflower oil, ethanol and water. APCBEE procedia, 5, 486-490.

Sherafatizangeneh, M., Farshadfar, C., Mojahed, N., Noorbakhsh, A., & Ardalan, N. (2022). Blockage of the Monoamine Oxidase by a Natural Compound to Overcome Parkinson's Disease via Computational Biology. Journal of Computational Biophysics and Chemistry, 21(3), 373-387.

Shunina Yu.S., Alekseeva V.A., Klyachkin V.N. (2015). Performance criteria for classifiers. Bulletin of Ulyanovsk State Technical University, 2(70), 67-70.

Tikhomirov, N. P., Tikhomirova, T. M. (2021). Methods of justification of effective demographic policy. In: Kitova, O., Dyakonova L. (eds.) Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020), Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov, 10, 52–62. Plekhanov Russian University of Economic, Moscow.

Tikhomirova, T. M., Sukiasyan A. G. (2020). Econometrics and Modelling in Management. Plekhanov Russian University of Economic, Moscow.

Tikhomirova, T. M., Sukiasyan A. G. (2021). Comparative estimates of human potential taking into consideration the risks of social inequality. In: Kitova, O., Dyakonova L. (eds.) Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020), Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov, 10, 63–76. Plekhanov Russian University of Economic, Moscow.

Tomek, I. (2010). Two modifications of CNN," In Systems, Man, and Cybernetics. IEEE Transactions on, 6, 769-772.

Voronkova, O. Y., Volokhova, T. V., Lebedeva, E. S., Smirnova, A. V., & Tubalets, A. A. (2022). Priorities for the development of medicinal plant growing in a post-pandemic environment. Siberian Journal of Life Sciences and Agriculture, 14(1), 436-451. doi:10.12731/2658-6649-2022-14-1-436-451