Building Models for Predicting Mortality after Myocardial Infarction in Conditions of Unbalanced Classes, Including the Influence of Weather Conditions

Irina Kashirina[0000-0002-8664-9817] and Mariya Firyulina[0000-0003-3468-5514]

Voronezh State University, 1 Universitetskaya pl., Voronezh, 394006, Russia
mashafiryulina@mail.ru

Abstract. The key direction in modern medicine is the development of software systems that allow us to analyze a large amount of data in an adaptive way and interpret the results obtained, ensuring high accuracy of results. Predicting mortality from myocardial infarction (MI) and identifying significant factors influencing this mortality is an urgent task, since the share of cardiovascular diseases annually accounts for more deaths than any other cause. The purpose of this study is to develop a machine learning model based on the gradient boosting technique for predicting mortality from MI, identifying the most significant signs, assessing the influence of meteorological factors, and improving the accuracy of predicting using data balancing methods. The paper used a depersonalized sample of all residents of the Voronezh region who had a STROKE in 2015-2017. The archive of weather data for 2015-2017 was taken from the site rp5.ru. Data analysis and model development were performed in the Python programming language version 3.6. Five undersampling strategies were reviewed and the method that achieves the highest prediction accuracy was chosen – the cluster centroid method.

Keywords: Undersampling, Gradient boosting, Predicting mortality, Myocardial infarction.

1 Introduction

Myocardial infarction (MI) is the most common and most fatal problem [1]. In medicine, information technologies are being actively introduced, which allow both to facilitate the work of staff and to reduce the mortality rate of the population. Enough software systems have been developed to predict diseases, reduce the risk of deaths, and select an effective method of treating patients depending on the nature of the disease. During solving such problems, the greatest attention is paid to improving the accuracy of the model and optimizing the software product. Improving the accuracy of the developed model can be achieved by more thorough pre-processing of data, or by considering additional factors that affect the result of predicting [2].

Machine learning algorithms are often referred to as black box models because there is no clear explanation of how exactly the model comes to predictions. Medicine is an industry where the interpretation of results is an important step. When determining signs that affect the risk of disease, you can eliminate these factors in a timely manner with the help of targeted medical intervention. Many works are devoted to identifying the dependence of the morbidity and mortality of the population on climatic conditions [3,4]. However, the results presented in these studies are inconsistent. The dependence of the functioning of the cardiovascular system on seasonality, fluctuations in the average daily temperature, humidity or pressure differ in different territorial zones, which is facilitated by the geographical location of the region, air pollution, the degree of adaptation of the population to the natural conditions of the place of residence, and so on. In order to check the presence of such dependences according to the data of the Voronezh region, climatic indicators were added to the clinical and medical indicators [5].

Thus, the purpose of this study was to construct models for predicting one-year mortality among residents of the Voronezh region who underwent MI, based on meteorological, socio-demographic, and clinical factors. The initial data were provided by the Voronezh Regional Cardiology Dispensary.

To build the model, one of the popular methods was chosen – gradient boosting, which is highly accurate [6].

In recent years, many works have been published devoted to predicting mortality from cardiovascular diseases, including those using gradient boosting methods [7, 8, 9]. The difference in this study lies in the addition of climatic indicators, as well as in a practical study of the effectiveness of class balancing methods in relation to the available source data.

During processing real data, there are often situations when the share of examples of one class in the training dataset is too small (minority class), and another is strongly represented (a majority class). Depending on the task, the problem of class imbalance can lead to a serious bias towards the majority class, a decrease in predicting efficiency and an increase in the number of false predictions. In tasks of a medical nature, this is especially important. When analyzing diagnostic results, there is often a problem of uneven distribution of classes in the training sample, since the number of patients suffering from a particular disease in the initial data is significantly lower (or, conversely, higher) than the number of healthy patients. Misclassification of the patient may lead to a lack of timely action. One approach to solve this problem is to use various sampling strategies. There are two ways to restore class balance. In the first case, delete a certain number of examples of the majority class (undersampling), in the second – increase the number of examples of the minority class (oversampling). This article discusses various undersampling methods and compares their accuracy.

2 Materials and Methods

2.1 Description of Source Data

For the analysis, we used depersonalized data on all patients who were admitted to Voronezh region’s hospitals in 2015-2017 with a diagnosis of MI, and dependencies in the sample with fatal cases of myocardial infarction (MI) for the same years. Total recorded cases of MI in 2015 – 3810, 2016 – 3837, 2017 – 3679. In 2015, 2016 and 2017, there were 684, 657 and 606 fatal cases of MI, respectively.

The source file contained information on 15 attributes presented in Table 1. The analyzed dataset was supplemented with six meteorological indicators. Weather data were downloaded from the rp5.ru website archive. Data preprocessing was performed using the Oracle SQL developer integrated development environment in the SQL language.

Table 1. Initial attributes.

Attribute's type	Attribute
Categorical variable	Gender, whether the myocardial infarction is repeated (MI), localization, KILLIP class, whether the patient underwent thrombolytic therapy (TLT), percutaneous coronary intervention (PCI), whether the patient has a history of diabetes mellitus (DM), atrial fibrillation (AF), acute cerebral circulation disorder (CCD), chronic obstructive pulmonary disease (COPD), chronic cardiovascular failure (CHF), arterial hypertension (AH)
Continuous variable	Age, Maximum of the temperature (Max_T), humidity, atmospheric pressure, wind speed, cloudiness

Table 2 shows the distribution of different values of certain characteristics among deceased and surviving patients. In article [10], the authors analyzed the influence of these factors on survival after MI using the Kaplan-Meier method, as a result of which it was noted that all predictors are important, except for gender and the patient's history of arterial hypertension.

2.2 Balancing Classes

When working with a real sample of data, there is often a situation often when the volume of one class are much larger than the volume of another. This data is called unbalanced. Building a model in such a situation may turn out to be ineffective and the error of the predicted data will be large. For medical issues, the data balancing stage is important. The predominance of the number of cases of one class leads to a shift in the model towards the majority class. Mortality prediction is a type of "early warning" tasks of an event. It is more important to predict the onset of death than to assume that the patient will survive and get the opposite result. Since the number of patients who died in the initial sample is much less than the number of survivors, it is necessary to balance the training data to correctly build a predictive model and reduce false negative results.

The class with the largest number of cases is called majority, with the smallest - minority. The class of the deceased in the problem under consideration is the minority class. Various sampling strategies are used to correct the class imbalance [11]. Rebalancing can be done in two ways: undersampling and oversampling. Oversampling is a duplication of minority class examples. Depending on what ratio of classes is needed, the number of random records for duplication is selected. Duplication occurs according to certain algorithms, for example SMOTE or ASMO. Undersampling is the removal of majority class examples.

Table 2. Summary of initial data.

	Surviving	Deceased
Killip class (%)		
I	51.6	22.33
II	30.28	23.27
III	10.89	19.98
VI	2.03	29.48
Age	65 ± 12	72 ± 12
PCI (%)		
Yes	8.81	2.46
No	91.19	97.54
CHF (%)		
H I	12.46	4.34
H IIA	38.84	41.55
H IIБ	2.04	9.21
H III	0.22	1.12
No	46.45	43.79
CCD (%)		
Ischemic stroke	3.79	8.6
TIA	0.46	0.61
Hemorrhagic stroke	0.23	0.36
No	95.53	90.43
COPD (%)		
Yes	6.58	11.96
No	93.42	88.04
Cloudiness	64±30	62±30
Max_T	11±13	12±13

Undersampling is considered the simplest and at the same time the most correct in the tasks of medical research. Therefore, it was this method that was chosen to solve the problem. The undersampling technique can be implemented in several ways. To achieve the highest prediction accuracy, five algorithms were considered, and the accuracy of their work was compared.

Algorithm for randomly deleting examples. To achieve the required class ratio, instances of the majority class are randomly deleted. The number of deleted records is determined empirically. The advantages of this method include simplicity of implementation and high performance. The main drawback is the high probability of losing significant data.

The method of cluster centroids. This strategy removes examples of the majority class using cluster analysis methods. The majority class is divided into the number of classes according to the number of instances of the minority class by the k-average method. Then the centroids of each class are selected, which eventually form a new class. The main advantage of this approach is that it preserves important topological properties of the sample.

Strategy for searching for Tomek links. All entries of the majority class that are included in the Tomek link are considered "noisy" and deleted. A pair *(E_i, E_j)* is called a Tomek link if there is no example of E_l such that the set of inequalities is valid:

$$\begin{cases} d(Ei, El) < d(Ei, Ej) \\ d(Ej, El) < d(Ei, Ej) \end{cases} \quad (1)$$

d – is the distance between the samples E_i, E_j.

The concentrated nearest neighbor rule. All examples are classified by the rule of the nearest neighbor. It then removes any instance whose class label differs from the class in at least two of its three nearest neighbors. The idea of this method is to remove instances from a majority class that is near or around the boundary of another class in order to improve the accuracy of classifying minority instances rather than majority instances [12].

The general parameter of XGBoost. This is a built-in data balancing method in the XGBoost – scale_pos_weight gradient boosting method, which allows you to increase the weight of the minority class examples.

2.3 Gradient Boosting and Quality Metrics

The gradient boosting machine learning model was used to predict the mortality of patients from MI. Five-fold cross-validation was used to train the model to correct hyperparameters, the final testing was carried out on a deferred validation set, the volume of which is 20% of the initial data.

Gradient boosting is a machine learning technique, the main idea of which is the iterative process of sequentially building partial models. Each new model is trained using information about the errors made at the previous stages, and the resulting function is a linear combination of the entire ensemble of models, considering the minimization of some penalty function [13]. This approach improves the generalizing ability and stability of the classification. This algorithm is distinguished by its high accuracy, which in most cases exceeds the accuracy of other methods.

One of the advantages of this method is its high performance, which is important when working with large data sets. The gradient boosting method is resistant to outliers, which are normal when working with real data.

In machine learning tasks, various metrics are used to assess the quality of models. When calculating these metrics, a classification error matrix is used. Based on the matrix elements for each class, the following indicators are calculated: (TP) – True Positive (the number of correctly predicted samples of class 1), (FN) – False Negative (the number of false negative samples, incorrectly predicted class 0), (TN) – True Negative (correctly predicted class 0), (FP) – False Positive (number of false positives, incorrectly predicted class 1). The main metric of classification problems is the proportion of correct answers, which is calculated using the formula:

$$accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

However, this indicator is ineffective for unbalanced samples. More often, the indicators are considered precision (precision) – the proportion of true positive examples from the total number of positively predicted examples, and recall – the proportion of true positive examples from the total number of positive examples.

$$precision = TP/(TP+FP) \tag{3}$$

$$recall = TP/(TP+FN) \tag{4}$$

Recall is preferable when evaluating models in medical problems. This metric helps to reduce the number of false negative predictions. Medical diagnostics has its own metrics that determine the accuracy of the method [14]. Sensitivity (true positive proportion) reflects the proportion of positive results that are correctly identified as such and the ratio of the number of deaths classified as deaths to the total number of deaths is calculated. If the deceased is considered a positive class, this metric coincides with recall. Specificity (true negative proportion) reflects the proportion of negative results that are correctly identified as such and the ratio of the number of survivors classified as survivors to the total number of survivors is calculated:

$$Specificity= TN/(TN+FP) \tag{5}$$

To create an optimal diagnostic system, it is necessary to find a compromise between the obtained indicators of sensitivity and specificity of the models. A common way to visualize the relationship between these metrics is to use the ROC-curve – a graphical characteristic of the quality of a binary classifier, the dependence of the sensitivity on the indicator (1-specificity) when varying the threshold of the decision rule of the model. The optimal position for the ROC curve is as close as possible to the upper left corner where specificity and sensitivity are at their maximum. The value of AUC ROC – the area under the ROC-curve is a compromise metric widely used in medical research [11].

3 Results and Discussion

Building a machine learning model and data analysis was performed using the libraries of the Python programming language. The gradient boosting model was built using the tools of XGBClassifier library. XGBoost is an optimized distributed library that is highly efficient, flexible and portable: you can train many models with different loss functions. The developed code is applicable in the main common environments (Kubernetes, Hadoop, SGE, MPI, Task).

Table 3 presents the metrics of the quality of the model's assessment, which were obtained based on true and false answers on the test sample: sensitivity, specificity, percentage of correct answers, and shows the percentage of correct answers in the final cross-validation. Table 4 shows the most significant features. The significance of each factor is calculated as the average normalized result of the decrease in the branching criterion caused by this factor. The branching criterion calculates the measure of uncertainty at the nodes of the trees. The Gini index was used as such a criterion.

Table 3. Model quality metrics gradient boosting.

Metrics	Values
Precision	0.78
Sensitivity	0.35
Specificity	0.97
The proportion of correct answers, Accuracy	0.85

Table 4. Feature importance.

Feature	Weight	Feature	Weight
Killip class	0.32	CCD	0.04
Age	0.11	COPD	0.04
PCI	0.08	Cloudiness	0.03
CHF	0.05	Max_T	0.03

The most significant feature is index of heart failure severity on the Killip scale. This is the expected result, since this indicator is determined by specialists when the patient is admitted to the hospital and characterizes the severity of the patient's condition during the initial visual examination. Age is expected to be a significant factor; survival is worse in patients belonging to the older age group. Also, predictors influencing the results include indicators: whether percutaneous coronary interventions (PCI) were performed, whether the patient has chronic heart failure (CHF), has a history of stroke (CCD) and chronic obstructive pulmonary disease (COPD). It can be noted that in comparison with clinical indicators, meteorological factors are less significant, however, there is an influence of cloudiness indicators (Cloudiness) and maximum daily temperature (Max_T).

As noted earlier, the initial sample has an imbalance in the data for the predicted variable – there are much more patients who survived after MI than who died. Therefore, despite the rather high indicators of Accuracy and AUC ROC, in Table 3 it can be seen that the constructed model has an extremely low sensitivity index (Sensitivity), which is unacceptable for the task of predicting mortality, since the model is poorly able to identify patients with a high risk of mortality.

An undersampling method is used to balance the classes, which can lead to an increase in the sensitivity of the machine learning model. The following undersampling strategies were considered: the random deletion of examples, the cluster centroid method, the Tomek link search strategy, and the concentrated nearest neighbor rule. Also considered is the built-in method of balancing data in the XGBoost gradient boosting method – scale_pos_weight.

Table 5 shows the main metrics for assessing the quality of the model: Accuracy (train) – the accuracy of the resulting model on the training set, sensitivity, specificity, Accuracy (test) – the proportion of correct answers on the test set, the area of the ROC-curve. Based on the results of the model quality metrics, we can conclude that the best indicators of accuracy when using strategies of random deletion, cluster centroids and Tomek links. In this case, the method of cluster centroids provides the highest sensitivity.

Table 5. Quality metrics for various sampling strategies.

Method	Accuracy (train)	Sensi-tivity	Speci-ficity	Accuracy (test)	AUC_R OC
Random undersampling	0.79	0.50	0.91	0.83	0.81
Cluster centriods undersampling	0.85	0.68	0.75	0.74	0.79
TomekLinks undersampling	0.85	0.40	0.98	0.86	0.82
NearMiss undersampling	0.80	0.48	0.88	0.81	0.76
Scale pos weight	0.83	0.16	0.99	0.83	0.80

Figure 1 shows a plot of ROC-curves for a gradient boosting model using various sampling methods. The best AUC ROC is 0.82.

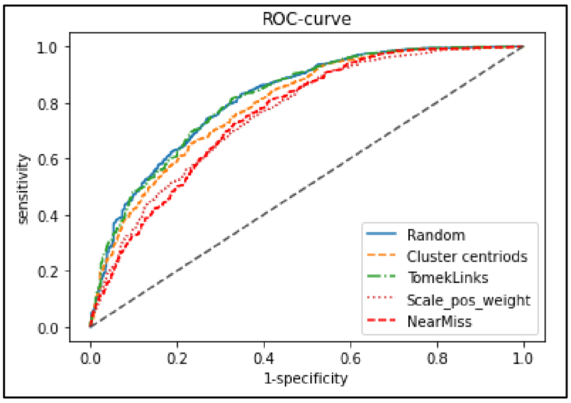

Fig. 1. Graph of the ROC curve of different sampling methods.

To assess the impact of weather conditions, the same models were built for a dataset without weather attributes. Table 6 presents the results. Comparing the indicators in Table 5 and Table 6, we can conclude that considering weather conditions helps to improve the accuracy of the model and has an impact on predicting mortality from MI.

Table 6. Quality metrics for various sampling strategies without weather predictors.

Method	Accuracy (train)	Sensi-tivity	Speci-ficity	Accuracy (test)	AUC_R OC
Random undersampling	0.74	0.28	0.99	0.80	0.81
Cluster centriods undersampling	0.80	0.30	0.99	0.84	0.77
TomekLinks undersampling	0.85	0.30	0.99	0.84	0.82
NearMiss undersampling	0.79	0.41	0.90	0.80	0.74

Table 7 shows the distribution of the number of patients in the test and training sets: the initial data and distribution after applying sampling methods on the training set, the test sample remains the original.

Table 7. Distribution of classes counts.

	Deceased	Surviving
All (Training)	2219	9581
TomekLinks undersampling	1359	8721
Random undersampling	2219	4500
Cluster centroids undersampling	2219	4700
NearMiss undersampling	2219	4700

4 Conclusion

This study was conducted to build a gradient boosting model for predicting mortality after myocardial infarction and to determine the most significant factors for mortality in MI. The accuracy of the model was improved by balancing the original sample using undersampling. The accuracy of five sampling strategies has been demonstrated. The most effective methods are the method of random removal of samples, the method of cluster centroids and the method of Tomek links. The best accuracy (Accuracy) was obtained using the Tomek links method (on test data, it is equal to 0.85). The best sensitivity is provided by the cluster centroid method.

With the help of the constructed model, significant factors were found for predicting mortality after the onset of myocardial infarction. The most significant are: Killip class, age, percutaneous coronary interventions, whether the patient has chronic heart failure, whether there is a history of stroke and chronic obstructive pulmonary disease, as well as weather factors – cloudiness and maximum daily temperature.

Acknowledgments. The research was supported by the Russian Foundation for Basic Research (RFBR), project № 20-37-90029.

References

1. Heron, M.: Deaths: Leading causes for 2017. National Vital Statistics Reports (6), 1–96 (2018)
2. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer (2013)
3. Dilaveris, P.: Climate Impacts on Myocardial infarction deaths in the Athens territory: the climate study. Heart (British Cardiac Society) 92(12):1747–1751 (2008)
4. Schwartz, J., Samet, J. M., Patz, J. A.: Hospital admissions for heart disease: the effects of temperature and humidity. Epidemiology (15):755–761 (2004)
5. Firyulina, M.: Influence of climatic conditions on mortality from myocardial infarction in the Voronezh region for 2015-2017. In Computer science: problems, methodology, technologies, pp 1256–1261. Publishing house “Research publications”, Voronezh (2019)
6. Firyulina, M., Kashirina, I.: Classification of cardiac arrhythmia using machine learning techniques. In: Journal of Physics: Conference Series. Actual problems of applied mathematics, computer science and mechanics 2019, vol. 10, pp. 1167–1175. Voronezh, Russian Federation (2019)
7. Weng, S. F., Reps, J., Kai, J.: Can machine-learning improve cardiovascular risk prediction using routine clinical data. PLoS One. 12(4) (2017)

8. Ambale-Venkatesh, B., Yang, X., Wu, C. O.: Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 121(9) (2017)
9. Garcia S., Herrera F.: Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. Evolutionary Computation 17(3), 275–306 (2009)
10. Kashirina, I., Firyulina, M., Gafanovich, E.: Analysis of the significance of predictors of survival after myocardial infarction using the Kaplan-Meyer method. In: Modeling, optimization and information technologies (24), pp. 7–20. Voronezh (2019)
11. Fernández, A., García, S., Galar, M.: Learning from Imbalanced Data Sets. 1st edn. Springer, New York (2018)
12. Brownlee, J.: XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. 2nd edn. Machine Learning Mastery Pty. Ltd, (2018)
13. Glazkova, T.: Assessment of the quality of diagnostic methods and prognosis in medicine. Bulletin of science center of medical Sciences of Russia (2), 3–11 (1994)
14. Shitikov V. K., Mastitsky S. E. Classification, regression, data Mining algorithms using R, https://ranalytics.github.io/data-mining/ (2017)