Computer Science Journal of Moldova, vol.**27**, no.1(79), 2019

Predicting the occurrence of strokes using the language R*

Vladimir Popukaylo

Abstract

Probability of stroke is analyzed based on data from Stroke.md system. Selection of features that significantly affect target variable was made. To solve the problem, classification algorithms are used: support vector machines, logistic regression, decision trees, random forest and others; comparison of the resulting mathematical models; preprocessing and building models produced in R-language.

Keywords: mathematical modeling, stroke prediction, data analysis, Stroke.md

MSC 2010: 68R10, 68Q25, 05C35, 05C05.

1 Introduction

This paper solves the problem of predicting stroke occurrence, based on data obtained from Stroke.md information system [1].

At the time of the research, the database contained information about 338 patients, including 96 patients who underwent routine medical examination and 242 patients who were hospitalized with stroke. These data were received while working on the State Program Risk factors, optimizing healthcare service, sustainable assessment and mathematical modeling of stroke [2].

The database contains information about 85 factors, including: personal data, lifestyle information, history, various types of laboratory tests.

* This work was supported by National Agency of Research and Development (NARD) project Ref. Nr. 17.000418.80.07A

To solve this problem, the R language [3] is used with libraries that implement various stages of data preprocessing, building models of classification and visualization of the results obtained. The main stages are implemented using the caret library, which implements a standard approach to machine learning [4].

2 Research problem statement

The research problem consists in creation of predictive model which will allow predicting with a certain share of confidence the probability of developing a stroke at a patient by results of collecting the anamnesis and carrying out some laboratory researches.

To carry out this study, a programming language R was chosen as a language specially created for conducting scientific researches and possessing a large number of additional libraries that facilitate both the preliminary analysis of data and the construction of predictive models.

When forecasting emergence of a stroke at the person it is necessary to solve a problem of classification which in our case is binary. Predictors in the constructed model should be some factors that are stored in the Stroke.md database and significantly affect the probability of stroke.

The solution of the problem can be divided into several stages:

1. Data conversion. At this stage, it is necessary to single out the target variable, to transform some factors into a form convenient for further automated processing, and also to remove non-variable and weakly filled factors from the dataset.

2. Selection of the signs which significantly influence the probability of developing a stroke.

3. Creation of classification models for unbalanced groups and assessment of their efficiency.

3 Data preprocessing

Data for research represent the database dump storing information on the patients who underwent preventive inspection (96 people); and patients who at the time of the research had a stroke in acute phase, divided into two waves of investigation (137 and 105 people).

Thus, the first step was to select the target variable, for which a vector variable was created that stores the value 1 if the patient had a history of any type of stroke before, and also if he had a stroke at the time of the study. At the same time, it turned out that 4 patients who had undergone a preventive examination did not have earlier information whether there had been a stroke, both ischemic and hemorrhagic. It was decided to reject these lines as not-informative. Also there were rejected not-informative columns, such as patients counting number, number of medical card and name of the table from which information was obtained. When analyzing the resulting table, it turned out that there are columns which are filled for less, than 30%. These are:

- Beta-lipoprotein (125 not available definition, hereinafter referred to as NAs).

- Thrombin clotting time (230 NAs).

- Years that a person has not smoked (contains 278 NAs), NAs are replaced by zeros.

- Group of obstetrical background (min. 125 NAs), NAs are replaced by zeros. This is due to the fact that this factor is not filled especially for men.

Thus, two more factors were excluded from consideration.

The following step became reduction of types for factors of initial data set: they were divided into quantitative and qualitative. R is a language with dynamic type-checking, but it does not always correctly recognize the type of a variable, therefore this procedure must be carried out manually. In addition, for qualitative variables, in the case of gradation of factor describing an unknown parameter value, it was replaced by NA. After carrying out these transformations, the parameters were rechecked for missing values and the columns filled for less than 30% were rejected, such as:

- Unruptured aneurysm.
- Metabolic syndrome.
- Valvulopathies.
- Acute myocardial infarction.
- Peripheral vascular disease in family history.

When constructing mathematical models that are based on various forms of data randomization, one has to deal with the problem of missing values. There are three main approaches to its solution:

1. Delete all incomplete lines. In our case, using this approach, we have to reject 244 lines, which is more than half of the available data and it does not seem to be rational.

2. Fill in the missing values using sample statistics of the corresponding variable, assuming that there is no interconnection between the variables. As we cannot suppose that the considered data is distributed according to the normal law, it is preferable to use the median for replacing the missing observations. Such a replacement ignores the specific characteristics of each patient, but it may not be a bad first approximation.

3. To fill the passed values taking into account linear correlation or one of proximity measures. As an example of such approach, the method of the closest neighbors, or intermediate models of regression and classification can be accepted.

The last step in data preprocessing was the removal of columns that store information about previously suffered strokes and near zero variance columns (Ruptured aneurysm, drug related).

After the conversion was completed, a table was obtained containing 334 rows and 71 columns, 30 of which are quantitative and 41 are qualitative.

Thus, an array of initial data was obtained, suitable for further processing by machine learning methods.

4 Feature selection

For selection of features that significantly affect the probability of stroke developing, classical statistical tests were carried out. Thus, for qualita-

tive data, Fisher's two-sided exact test was used as the most powerful of the existing methods, and the Wilcoxon-Mann-Whitney test was used to analyze quantitative data, since the hypothesis about the normal distribution of factors in dataset was not tested. To solve the problem of multiple comparisons, two approaches were used: control of group error probability of the first kind (Bonferroni correction) and control of the average proportion of false deviations (Benjamin-Hochberg method).

This approach allowed us to identify the following factors as significantly influencing the likelihood of stroke (using the Bonferroni amendment):

Qualitative factors are:

- Fasting ($p < 0.0000001$).
- Tooth extraction ($p < 0{,}0000001$).
- Hypertension ($p = 0.00002$).
- Rhythm ($p = 0{,}00034$).
- Marital status ($p = 0.00054$).
- Paradontosis ($p = 0.00383$).
- Occupation ($p = 0.01205$).
- Systemic diseases ($p = 0{,}01237$).
- Atrial fibrillation ($p = 0.01698$).
- Physical activity ($p = 0.02190$).

Quantitative factors are:

- Age ($p < 0.0000001$).
- Pregnancies ($p < 0.0000001$).
- Fibrinogen ($p < 0.0000001$).
- Systolic Blood Pressure ($p = 0.000002$).
- Glucose ($p = 0.00282$).
- Spontaneous abortions ($p = 0.01904$).
- Diastolic Blood Pressure ($p = 0.02101$).
- Leukocytes WBC ($p = 0.03710$).

Usage of less conservative amendment of Benjamin-Hochberg allowed identifying the following parameters in addition: Births; Activated partial thromboplastin time APTT; Sleep Disorders; Address; Sex; Old myocardioal infarction; Migraine.

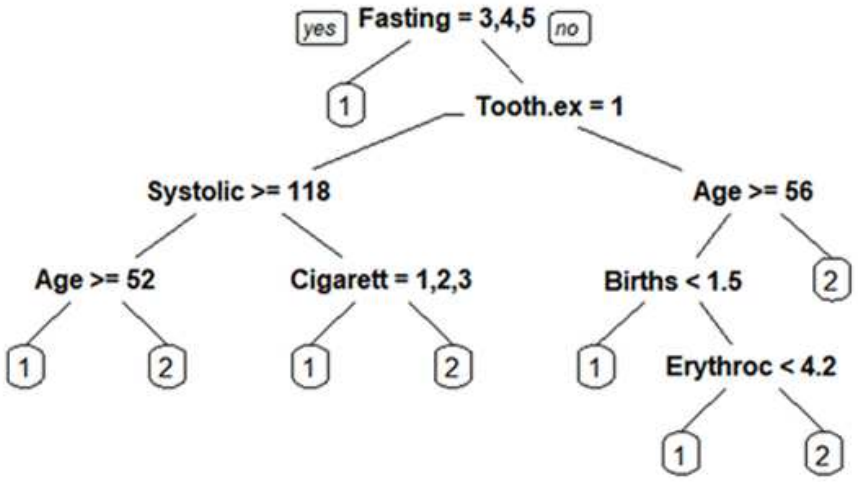

Figure 1. Decision Tree

An alternative method of searching for significant factors is to build a decision tree. So, using the 10-fold cq-cross validation, the following most significant parameters were selected: Fasting, Age, Tooth extraction, Systolic blood pressure, Fibrinogen, Births, Hemoglobin, Activated partial thromboplastin time APTT, Pregnancies, Sex, Dental prosthesis, Creatinine, Paradontosis, Ocupation, Height, Cigarette smoking, Hypertension, Antihypertensive, Medical abortions, Hematocrit HCT, Erythrocytes RBC, ASAT, Diastolic blood pressure, Marital status, Leukocytes WBC, Cholesterol.

Figure 1 shows the image of one of the constructed trees.

The considered algorithms of searching the factors which are significantly influencing a target variable allow identifying a number of the factors requiring closer attention at prevention of strokes.

It should be noted that the list of factors selected by each of the methods is somewhat different from that generally accepted in the lit-

erature [5] on this topic.

This could depend on the characteristics of the data collected. However, our task was to develop a model based on the available information, so it was decided to use the above columns as predictors.

5 Creation of predictive models

For training the models and selecting their hyperparameters, the caret library was used, which provides universal interface for training models of various types.

At the stage of selecting the structure of the model, a 10-fold repeated cross validation was carried out. Different types of models have been tested, which are: logistic regression, decision trees, support vector machine and various boosting algorithms. Figure 2 shows that algorithms based on the construction of model ensembles show more reliable results, both in accuracy prediction and in agreement between classes.

The best results were obtained using such models as: Adaboost.M1; Random Forest; DeepBoost; Extreme Gradient Boosting on Trees; Boosted Generalized Linear Model; Extreme Gradient Boosting on Linear models.

Table 1 presents the numerical values for the Accuracy parameter, obtained by cross-validation for some of the algorithms.

Most of the known boosting models based on the base models of various types show statistically indistinguishable results with a median accuracy 0.8261. At the same time confidential intervals show insignificant advantage of such algorithms as Adaboost for decisions trees models and Extreme Gradient Boosting for linear models.

Comparison of coherence between classes on Kappa Kohen's criterion also confirms previous findings. At the same time, the average consistency of classes for the specified algorithms varies at a level of 0.5, which indicates satisfactory agreement.

It is known that adaptive boosting models can screen out insignificant parameters, however, in the case of supersaturated tables, the algorithms may not give the best indicators.

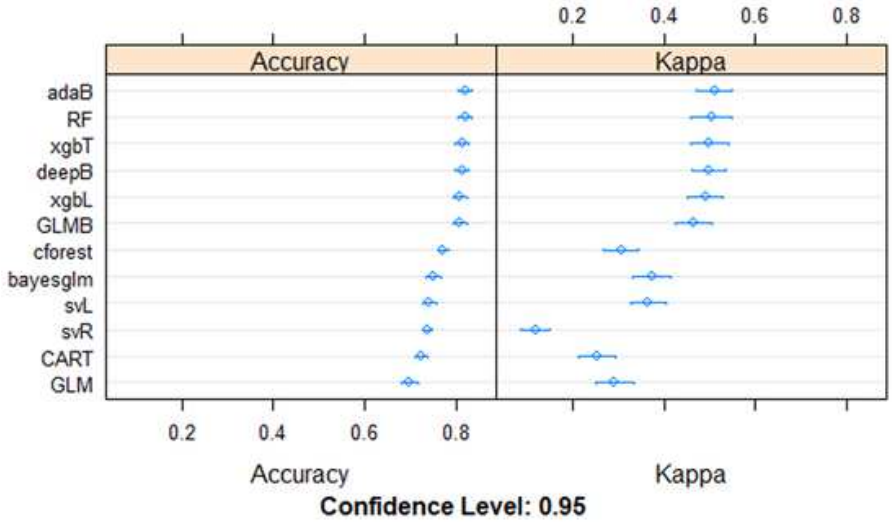

Figure 2. Accuracy and Kappa for models with all predictors

Table 1. Accuracy of models with all predictors

	$Min.$	$1st.Qu.$	$Median$	$Mean$	$3rd.Qu.$	$Max.$
$adaB$	0.6522	0.7826	0.8130	0.8193	0.8696	1.0000
$cforest$	0.5833	0.7391	0.7826	0.7719	0.8261	0.9130
$GLMB$	0.6667	0.7500	0.7917	0.8086	0.8696	0.9583
$BGLMB$	0.4800	0.6957	0.7500	0.7513	0.8261	0.9167
$deepB$	0.6522	0.7500	0.8261	0.8130	0.8696	0.9583
$CART$	0.5417	0.6667	0.7391	0.7243	0.7826	0.8750
RF	0.6087	0.7500	0.8261	0.8191	0.8750	1.0000
GLM	0.4783	0.6522	0.7083	0.6978	0.7500	0.8800
$svmL$	0.4167	0.6957	0.7391	0.7414	0.7917	0.9167
$svmR$	0.6087	0.7083	0.7391	0.7360	0.7826	0.8333
$xgbT$	0.6522	0.7500	0.8261	0.8132	0.8696	0.9583
$xgbL$	0.6250	0.7826	0.8261	0.8088	0.8696	0.9565

In this connection, it was decided to train the models with the most important selected parameters. The best results were obtained when selecting significant features for statistical tests with a significance level of p <0.05 and a Benjamin-Hochberg correction based on a false discovery rate.

The Accuracy values received on cross-validation for the algorithms which improved the indicators are presented in Table 2.

Table 2. Accuracy of models with 25 predictors

	$Min.$	$1st.Qu.$	$Median$	$Mean$	$3rd.Qu.$	$Max.$
$adaB$	0.6522	0.7826	0.8333	0.8292	0.8750	1.0000
$cforest$	0.6250	0.7391	0.7826	0.7863	0.8474	0.9130
$GLMB$	0.6522	0.7826	0.8261	0.8223	0.8696	0.9565
$BGLMB$	0.6087	0.7500	0.8130	0.8078	0.8696	1.0000
$deepB$	0.6522	0.7826	0.8333	0.8275	0.8750	0.9583
$CART$	0.5417	0.6957	0.7500	0.7483	0.7917	0.9200
RF	0.6957	0.7917	0.8261	0.8291	0.8696	0.9583
GLM	0.5833	0.7473	0.7917	0.7972	0.8696	0.9583
$svmL$	0.6250	0.7391	0.7917	0.7911	0.8333	0.9565
$svmR$	0.6522	0.7826	0.8261	0.8203	0.8696	1.0000
$xgbT$	0.6667	0.7917	0.8696	0.8521	0.9130	1.0000
$xgbL$	0.6522	0.7917	0.8514	0.8427	0.9130	1.0000

The confidence intervals for the Accuracy and Kappa parameters are presented in Figure 3.

Thus, it can be seen on Figure 3 that the best results at the stage of cross-qualification were shown by the Extreme Gradient Boosting on Tree algorithm (xgbT). This algorithm showed the best tune with hyper parameters:

- eta = 0.4;
- nrounds = 150;
- max depth = 3;
- colsample by tree = 0.6.

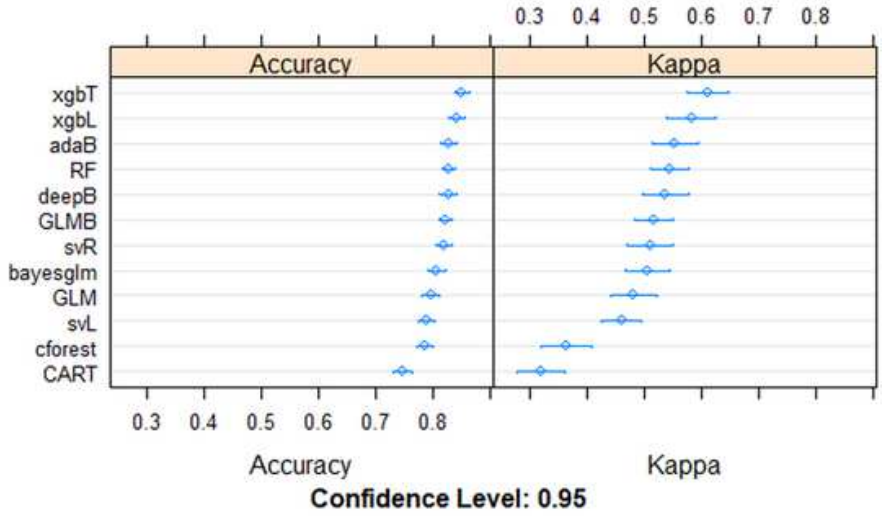

Figure 3. Accuracy and Kappa for models with 25 predictors

Among non-boosting models, the best results were obtained using support vector machine with radial kernels.

At the next stage, the accuracy of the selected algorithms for the holdout dataset (99 items) was evaluated. The results obtained for the models that showed the best results are presented in Table 3.

6 Conclusions

During the conducted research it was succeeded to improve the results received earlier [5]: to construct the predictive models capable with an approximate accuracy at the level of 80% to classify patients who had stroke; parameters which can influence the occurrence of a stroke were selected. The best quality in forecasting was reached by means of the algorithm of extreme gradient boosting on trees of decisions. Median accuracy for this algorithm at a stage of cross-validation equals 86.96%,

Table 3. Accuracy of models for the holdout dataset

Algorithm	*Accuracy*
Extreme Gradient Boosting on Decision Trees	0.77778
Random Forest	0.79798
Boosted Generalized Linear Model	0.78788
Extreme Gradient Boosting on Linear Model	0.78788
Support Vector Machine with Radial Kernel	0.79798
Support Vector Machine with Linear Kernel	0.75757
Generalized Linear Model	0.74747
Bayes Generalized Linear Model	0.77778
*AdaBoost.M*1	0.74747
DeepBoost	0.72727

on the holdout dataset accuracy equals 77.78%. Among the models which are not based on randomization, the best results were shown by support vector machine with radial kernel. Median accuracy of such model at a cross-validation stage – 82.61%, on the holdout dataset – 79.8%.

The results can be used in the software systems development targeted at preventing strokes among the population.

References

[1] E. Zamsa, "Medical software user interfaces, stroke MD application design," in *2015 E-Health and Bioengineering Conference – EHB*, 2015, pp. 1–4.

[2] S. Groppa, N. Ciobanu, D. Efremova,"Stroke risk factors in the population of Republic of Moldova," *Journal of the Neurological Sciences*, vol. 381, p. 411, Oct., 2017.

[3] M. Kuhn, "Building predictive models in R using the caret package," *Journal of statistical software*, vol. 28, no. 5, pp. 1–26, Nov., 2008.

[4] RC Team. *R language definition* (2019) [Online]. Available: ftp://155.232.191.133/cran/doc/manuals/r-devel/R-lang.pdf

[5] S. Cojocaru *et al.*, "Analysis and preparation of data from Stroke.md database when creating a stroke prediction model," in *Proc. MFOI-2018*, 2018, pp. 51–66.

Vladimir Popukaylo Received March 18, 2019

Vladimir Popukaylo
Vladimir Andrunakievich Institute of Mathematics and Computer Science
T.G. Shevchenko Transnistrian State University
Moldova, Tiraspol, str. Vosstania 2a, cab. 311B
Phone: +37377844001
E–mail: vsp.science@gmail.com