Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3013

ANALYSIS AND PREDICTION OF INFLUENCING FACTORS ON THE PROBABILITY OF STROKE

XinChun Wang
School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 2545083901@qq.com

Abstract: With the continuous development of big data, statistical model analysis has permeated various fields of life, particularly in clinical medicine. In addressing the clinical prediction of patients' stroke probability, modeling methods such as ANOVA, support vector machines, binomial logistic regression analysis and random forest models are essential. Given that the dependent variable in this study is a binary classification problem, logistic regression analysis and random forest models were selected for the analysis. This paper elaborates on the principles of logistic regression analysis and random forest models and random forest models, providing the regression equation for the regression model and the importance scores of each variable in the random forest model. Additionally, the predictive capabilities of these two models were evaluated, including an assessment of prediction accuracy.Through the application of regression and random forest models, we aim to enhance the clinical prediction accuracy of patients' stroke probability, thereby providing a more reliable basis for clinical decision-making.

Keywords: Stroke; Influencing factors; Logistic Regression Analysis; Random forest regression model; Prediction accuracyt

1 INTRODUCTION

With the increasing development of data information and data technology, the application field of statistics has become more extensive, especially in the field of clinical medicine. Clinical prediction models refer to establishing models using various relevant factors to calculate the probability of the occurrence of a certain disease or the future disease status[1]. Clinical risk prediction models mainly involve doctors using established predictive models based on various medical influencing factors to analyze and calculate the overall probability of the future occurrence of a specific disease or the risk of future disease for patients[2]. Diagnostic probability models focus on analyzing the probability of a particular clinical disease diagnosis in an individual patient, determining the likelihood of diagnosing a disease based on clinical signs and characteristics of the patient or a specific group, predicting the current status of the patient (whether they are ill), which is more common in cross-sectional studies[3]; prognostic models are used to predict the likelihood of an individual patient experiencing a specific event in the future, mainly referring to recurrence, death, disability, and future complications, generally in cohort studies[4].

Logistic regression models and random forest models are widely used in the field of clinical prediction. For example, factors influencing the occurrence of a certain disease are explored, and the probability of disease occurrence is predicted based on these factors[5]. In this study, taking the analysis of the probability of stroke as an example, two groups of people are selected, one group with stroke and the other without stroke, with different disease characteristics and daily lifestyles. Therefore, the dependent variable is whether the person has a stroke, with values of "yes" or "no," and there can be many independent variables, such as age, gender, whether they have heart disease, whether they have hypertension, etc[6]. Independent variables can be either continuous or categorical.

2 DATA SOURCES AND DESCRIPTIONS

The data in this article is from the stroke prediction dataset published on the website. The analysis of this study involved a total of 1110 research subjects. The original dataset had 11 variables, with the response variable being whether or not the individual had a stroke, which is a binary variable. The occurrence of stroke is influenced by multiple factors, including both quantitative and qualitative data. For the convenience of this study, the indicators affecting the occurrence of stroke are mainly represented by quantifiable variables(Table 1).

Table1 Variable Description Table

Symbols	Variable Name	Illustrate
Y	Stroke	Y=1:Yes; Y=0:No

A	Age	A=0.25:age in (0,31]; A=0.5:age in (31,52]
		A=0.75:age in (52,68]; A=1:age in (68,100]

G	Gender	G=1:Male; G=0:Female
X_1	Hypertension	X_1=1:Yes; X_1=0:No
X_2	Heart disease	X_2=1:Yes; X_2=0:No
X_3	Married	X_3=1:Yes; X_3=0:No
X_4	Work Type	X_4=1:employed; X_4=unemployed
X_5	Residence type	X_5=1:Urban; X_5=0:Rural

X_6	Avg glucose level	Numeric
X_7	Body Mass Index	Numeric

Before modeling, conduct descriptive statistics on the data. Among 1110 samples, 249 individuals suffered from stroke, with 108 males and 141 females(Figure 1). The likelihood of women suffering from a stroke is slightly higher than that of men.

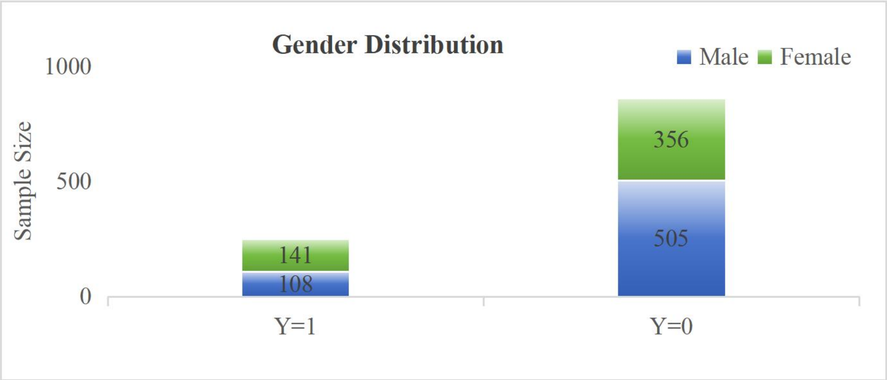

Figure 1 Gender Distribution

From the age distribution chart, it is evident that the largest number of samples falls within the older age group(Figure 2). The probability of stroke is highest among middle-aged individuals. However, it should not be overlooked that young adults also exhibit a certain incidence of stroke, indicating that advanced age is not the sole factor for stroke occurrence. This underscores the objectivity and importance of data analysis.

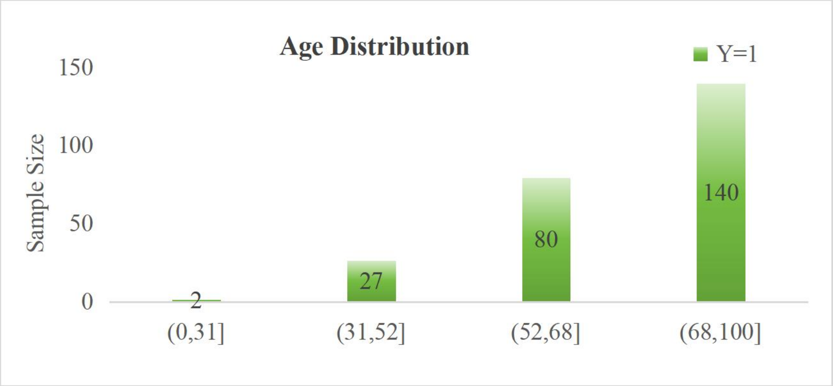

Figure 2 Age Distribution

Additionally, among other phenomena, the highest numbers of afflicted individuals are those who are married and employed, while the numbers of those with hypertension and heart disease are the lowest(Figure 3). However, it is important to note that this does not imply that stroke is unrelated to hypertension and heart disease. This is not only because descriptive statistics do not allow for direct conclusions, but also because hypertension and heart disease are merely factors that can trigger strokes; they are related but not prominently. Our study focuses on risk prediction for patients who have not yet had a stroke, analyzing the characteristics of relevant phenomena during the latency period.

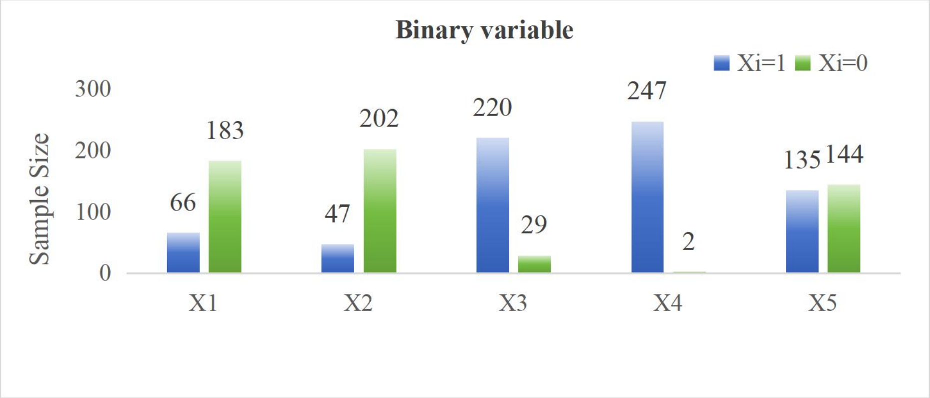

Figure 3 Binary variable Distribution

Next, we further explore the relationship between variables and stroke(Figure 4). By calculating the correlation coefficients between the independent variables and the dependent variable, as well as the correlation coefficients between each pair of variables, we can draw some conclusions. From the graph, we can see that the correlation coefficient between age and stroke is the highest, indicating that, the older the age, the higher the future risk of stroke. On the other hand, the correlation coefficients for gender, residential type, and body mass index are almost zero, suggesting that these three variables may not be related factors for stroke.

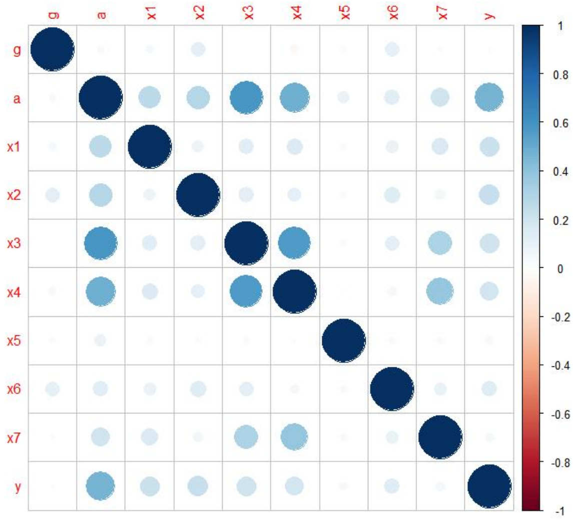

Figure 4 Correlation Coefficients

3 ANALYSIS OF FACTORS INFLUENCING STROKE INCIDENCE

There are many factors that contribute to stroke, such as heart disease and hypertension. Most of these indicators are categorical variables that do not follow a normal distribution and have high collinearity among explanatory variables. Therefore, traditional regression models do not meet the necessary assumptions. Consequently, traditional regression models are abandoned. In this section, a random forest model is constructed to evaluate the probability of stroke occurrence.

3.1 Implementation of Logistic Regression Models

The number of decision trees that corresponds to the minimum out-of-bag error is 328, determined by building a random forest model.

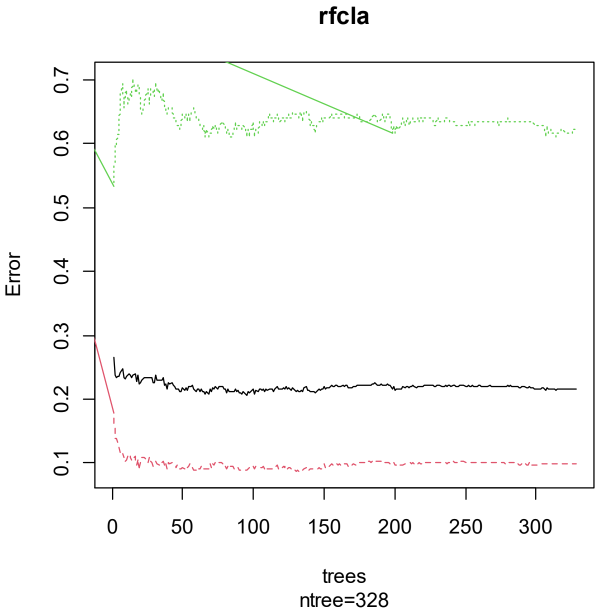

Figure 5 Out-of-bag Error/ntree=328

As shown in the above figure 5, this chart depicts the decreasing trend of out-of-bag data obtained from the random forest model. When the number of decision trees increases, the out-of-bag data tends to stabilize.From the diagram, it can be roughly estimated that the out-of-bag error rate is around 60%, with the Type I error rate and Type II error rate being approximately 60% and 10% respectively. Below, the specific values for the out-of-bag error rate, Type I error rate, and Type II error rate are calculated.

Next, we will visualize the out-of-bag error rate, the first type error rate, and the second type error rate, and calculate their specific values(Table 2).

Table2 Out-of-bag Error Rate and **Confusion matrix**

OOB estimate of error rate: 21.62%			
Confusion matrix:			
	0	1	class.error
0	543	59	0.09800664
1	109	66	0.62285714

The Type I error rate and Type II error rate are respectively at 62.3% and 9.8%, with the Type I error rate being somewhat high. However, the out-of-bag error rate is at 21.62% , indicating that the prediction accuracy of this random model is at 78.38%, which shows that the overall predictive performance of this random model is relatively good.

3.2 Implementation of Logistic Regression Models

In a random forest model, it is not possible to obtain the average regression coefficient for each independent variable individually. Instead, the importance of each variable should be determined using the comprehensive scoring method, which involves the average decrease in mean squared error and the average decrease in the accuracy error of the forest model. These scores evaluate the average impact of each variable on the weights of other dependent variables. Within the research framework of this paper, the influence and extent of each independent variable on the primary dependent variable are illustrated in the following diagram.

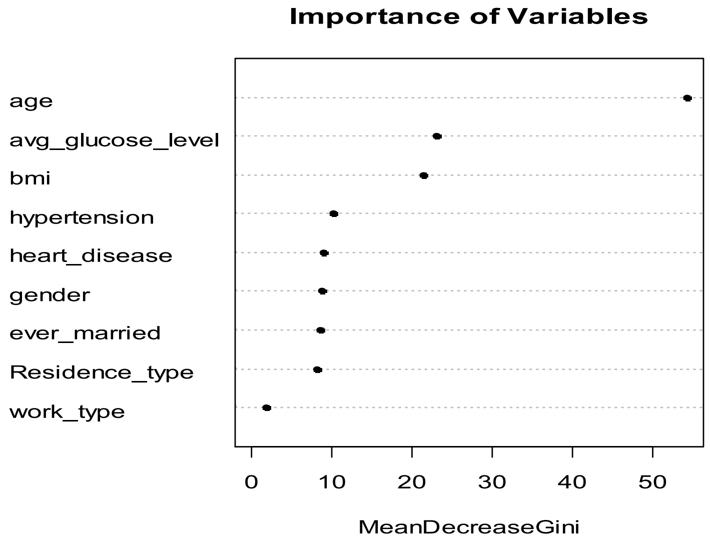

Figure 6 Importamce of Variables

In Figure 6, it can be observed that, according to the mean squared error values, the top five significant factors influencing stroke are age, average glucose level, body mass index, hypertension, and heart disease. Among these, age is the most significant factor affecting stroke, indicating that the older a person is, the higher the probability of having a stroke. This highlights that elderly individuals are particularly at risk of stroke. The second most significant factors are average glucose level and body mass index, suggesting that to reduce the probability of stroke, it is crucial to control the increase in average blood glucose level and body mass index by managing blood sugar levels and preventing obesity in daily life. Additionally, attention must be paid to the prevention and control of hypertension and heart disease. Although these two factors are not as critical as the first three, they are still contributors to stroke and should not be ignored. Patients with hypertension or heart disease must control the progression of these conditions to prevent inducing a stroke. Among the factors, although indicates that married individuals have a higher incidence rate than unmarried individuals, this might be because married individuals are generally older than unmarried ones, so it is not considered for now. Residence type has a certain impact on stroke, but its influence is not significant. Occupation work type is the least important factor affecting stroke in the random forest model, so having or not having a job has the least impact on stroke.

3.3 The Confusion Matrix of the Random Forest Model and Prediction Accuracy

Table3 The Confusion Matrix of the Random Forest Model

Actual Predict	0	1
0	230	28
1	43	31

Just like Table 3, using this confusion matrix, we can determine that the prediction accuracy of the random forest model is 78.6%, which is greater than 75%, indicating that this model has good prediction accuracy and strong predictive ability.

4 CONCLUSION

We use a random forest model to determine the number of decision trees that result in the smallest error. The first type of error rate and the second type of error rate are respectively, with the first type of error rate being a bit high, but the out-of-bag error rate is, indicating that the prediction accuracy of this random forest model is 78.38%, which means the overall prediction performance of this random forest model is quite good. Among the nine influencing factors, based on the mean squared error, the top five factors significantly impacting stroke are age, average glucose level, body mass index, hypertension, and heart disease. Among these, age is the most significant factor affecting stroke; the older the age, the higher the probability of having a stroke. This indicates that elderly people are at a particularly high risk of stroke. Secondly, average glucose level and body mass index have the highest impact on stroke, implying that controlling average glucose levels and body mass index is crucial in reducing the probability of stroke. It is important to manage blood sugar levels and prevent obesity in daily life. Lastly, attention should also be paid to the prevention and control of hypertension and heart disease. Although these two factors are not as critical as the previous three, they are still factors that can induce strokes and should not be ignored. Patients with hypertension or heart disease must control the progression of their conditions to prevent strokes. The prediction accuracy of this random forest model is 78.6%.

COMPETING INTERESTS

The author have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

[2] Liaw A., Wiener, M. Classification and regression by randomForest. R news, 2(3): 18-22.

[3] Diaz-Uriarte, R., Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 2006, 7(1): 3.

[4] Ishwaran H., Kogalur UB. Random forests for survival, regression and classification (RF-SRC). R packag e version 1, 2007, 4.

[5] Ishwaran H., Kogalur UB., Blackstone EH., Lauer MS. Random survivalforests. The annals of applied sta tistics, 2008, 2(3): 841-860.

[6] Genuer Robin, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using random forests. Pattern Recognition Letters, 2010, 31(14): 2225-2236.