Stroke Prediction Base on Logistic Regression Model

Le Li *

College of Natural science, University of Texas at Austin, Austin, Texas, TX 78705, United State

* Corresponding Author Email: laylali@utexas.edu

Abstract. The prediction of stroke using risk factors and basic demographic information can be valuable for primary prevention and community healthcare workers. Machine learning models are becoming more prevalent in clinical prediction due to their high accuracy. This study investigates stroke prediction using four machine learning models including logistic regression, random forest, Lasso Regression (Least Absolute Shrinkage and Selection Operator), and Extreme Gradient Boosting (XGBoost). Six key variables—age, gender, work type, heart disease history, marital status, residency type, smoking status, Body Mass Index (BMI), and average glucose level—are used for prediction. Among the dataset of 3,255 observations, 180 individuals experienced stroke, indicating an extremely imbalanced dataset. Therefore, balanced accuracy is the key metric used to compare model performance. The balanced accuracy for logistic regression, Lasso Regression, random forest, and XGBoost are 78.4%, 61.7%, 50%, and 67.7%, respectively. Logistic regression demonstrated the strongest performance, while also highlighting the significant role of age, particularly for individuals over 45.

Keywords: Stroke prediction, random forest, logistic regression, LASSO, XGBoost, Machine learning.

1. Introduction

Stroke is one of the leading cause of deaths and long-term disability worldwide. Stroke also increases the risk of another chronic disease, one in six deaths due to cardiovascular disease was due to stroke [1]. On the other hand, even patients are healed, still have a risky chance of permanent disability and incapacity, and reduced social activities. Goldstein classifies risk factors into three categories: nonmodifiable (demographic factors), well-established modifiable (such as blood pressure, lifestyle choices, and atrial fibrillation), and less well- established or potentially modifiable (including metabolic syndrome, alcoholism, and substance abuse) [2]. Addressing the Well-established and potentially modifiable risk factors can lower stroke risk, highlighting the importance of early detection and prevention for improving outcomes during the critical period.

Based on risk factors and well-collected medical record datasets, machine learning models are widely used in stroke prediction. Liu's research utilizes convolutional neural networks to extract features (including hypertension, heart rate, medical history, etc.) and combines these features through a neural network model, achieving 98.53% accuracy in predicting the probability of stroke [3]. The ASTRAL score model developed by Ntaios demonstrates nearly 85% accuracy in forecasting functional impact of acute ischemic stroke (AIS) using patient information collected in the emergency room, by applying a logistic regression model [4]. Heo implements machine learning models, involving random forests model and deep neural networks model, comparing them with Ntaios's ASTRAL score prediction, with respective accuracies of 88.8%, 85.7%, and 83.8% [5].

Previous research has primarily concentrated on medical check-ups or records, such as heart rates, atrial fibrillation, and other diagnostics. This study aims to create a relatively low-barrier model that does not require hospital screenings or advanced diagnostic tests, enabling potential patients, especially older individuals or those concerned about their health, to complete less costly and convenient. The goal is to provide a reference for early medical intervention and enhance awareness of health behaviors among individuals, families, and community health workers.

2. Method

2.1. Data and Variable

For the development of a machine learning model, this research utilizes the dataset from Kaggle was published in 2020 and includes a total of 3255 observations [6]. Variables included in the machine learning model are age which are classifies in 6 age groups 18-24, 25-34, 35-44, 45-54, 55-64 and 65 above, gender, work type, heart disease history, marital status, residency type, smoking status, bmi values and average glucose level. Individuals reported with strokes are 180 compare to 3075.

2.2. Machine Learning Algorithms

This paper demonstrates a comprehensive applications and analysis of four machine learning models, logistic regression, lasso regression, random forest, and XGBoost to forecasting the possibility of stroke.

Logistic Regression is a simple classification model with linear approach to forecast the likelihood of binary outcomes. Compared to other machine learning methods, logistic regression is tended to overfitting less, especially when variables are limited. Moreover, when variables are linearly relative to the outcomes, logistic regression demonstrate a strong performance. Lasso Regression (Least Absolute Shrinkage and Selection Operator) improves forecasting accuracy by incorporating both variable selection and regularization. It helps in selecting the most important features through shrinking the less significant ones toward zero, effectively reducing the risk of overfitting. Random Forest is an aggregate model with numerous decision trees established in training data. Each decision tree is considered as evidence for prediction, and the final forecasting is based on majority voting or average of the individual predictions. Result from this aggregation method is more robust and accurate than single decision tree model -- Classification and Regression Tree (CART). Extreme Gradient Boosting (XGBoost) is a boosting algorithm that iteratively minimizes the residuals or errors produced by the previous model in each step. It also includes a scale weight parameter that can be used to assign higher weights to positive examples, making it specifically powerful in processing imbalanced datasets. To ensure robust performance, K-fold cross-validation is applied for the logistic regression, lasso regression, and XGBoost models. This method allows for a systematic approach to assessing the models' accuracy by partitioning the dataset into K subsets. The most favorable positive probability threshold can be identified by training the model in different subsets, ultimately enhancing the reliability of the model's outputs. In random forest application, we employ 500 trees in the model and set mtry to 1, meaning that only one predictor variable (feature) is considered at each split, based on insights gained from 10-fold cross-validation.

The training population is composed of a substantial number of participants, with 80% (n=2604) randomly selected to form the data. This allows for extensive model training and tuning. The remaining 20% (n=652) of the participants are set aside as the test set, serving as an independent dataset to validate the models. This split is target for preventing overfitting, ensuring that the models generalize well to data.

3. Result

In 3255 observations, 180 data are missing in the smoking status and the age value under 1 which is considered as invalid. Among 3025 valid observations, average age of having stokes are 67.7, average bmi for individual with strokes is 30.47 which is considered as obesity.

3.1. Significant Variables

In the analysis of the ten variables, certain factors emerged as particularly significant in their association with stroke risk. Notably, hypertension, heart disease, and average glucose levels demonstrated a strong influence on stroke occurrence, each exhibiting a p-value whihch less than

0.01. This reveals a highly statistically significant relationship, suggesting that individuals with elevated blood pressure, a history of heart disease, or higher average glucose levels are at an increased risk for stroke. Conversely, the impact of smoking status was found to be relatively less significant, with an estimated p-value around 0.2 in the model. While smoking is a well-documented risk factor for numerous health issues, including stroke, the weaker association in this study suggests that its effects may be overshadowed by the more potent influences of hypertension and heart disease.

Additionally, the analysis revealed that age plays a crucial role in stroke predictions, particularly for individuals aged 45 and older. The three age groups within this range—45-54, 55-64, and 65 and above—exhibited a progressively increasing risk for stroke

3.2. Comparison of the Models

Due to the extreme unbalced response of stroke disease, 180 compare to 3075. The drawback of the measrues -- accuracy can be misleading. For example, in the random forest model, the accuracy is 0.94497 which is the higest among other 3 model, but the positive prediction is zero. In comparsion, F1 score focuses on the minority class performance and balanced accuracy which gives equal weight to both classes are better determined the unbalanced datasets. Thus, in evaluating the performance of four machine learning models—logistic regression, lasso regression, random forest, and XGBoost—we compared their balanced accuracy and F1 scores to understand their effectiveness in predicting stroke outcomes.

In confusion matrix demonstrated in table 1, the logistic regression model exhibited the highest F1 score of 0.2588 and a balanced accuracy of 0.78413. This reveals that although the model struggles with predicting positive observations, it maintains a reasonable overall balance between sensitivity and specificity. Lasso regression, with an F1 score of 0.1768 and balanced accuracy of 0.61734, reflects even greater challenges, suggesting poor performance in identifying positive instances. The random forest model, despite achieving a high balanced accuracy of 0.5, demonstrated an F1 score of 0, indicating a total failure to predict any positive cases. This highlights a significant issue: while the model has a preference in prediction negatives, it does not generalize well to the minority class. XGBoost showed similar challenges, with an F1 score of 0.1683 and a balanced accuracy of 0.67657, indicating that it more struggled to effectively classify positive than random forest.

Overall, the comparison of the F1 score and balanced accuracy across these models reveals important insights into their performance. While logistic regression reveals a best prediction with the highest balanced accuracy, its F1 score underscores the need for improvement in identifying positive cases. Random forest's high balanced accuracy is misleading, as it fails to predict positives altogether.

There are several limitations for this research, the unbalanced data is the major problems.

Table 1. Comparison confusion matrix

Measure	Logistic regression	Lasso regression	Random Forest	XGboost
Sensitivity	0.8684	0.4444	0	0.8947
Specificity	0.6998	0.7902	1	0.4584
Precision	0.1521	0.1103	N/A	0.0929
Negative Predictive Value	0.9885	0.9605	0.9447	0.986
False Positive Rate	0.3002	0.2098	0	0.5416
False Discovery Rate	0.8479	0.8897	N/A	0.9071
False Negative Rate	0.1316	0.5556	1	0.1053
Accuracy	0.7097	0.7711	0.9447	0.4839
F1 Score	0.2588	0.1768	0	0.1683
Balanced Accuracy	0.78413	0.61734	0.5	0.67657

4. Discussion

The results from these applications of machine learning to stroke prediction provide valuable interpretations, particularly in the context of imbalanced datasets. The significant variables identified, including hypertension, heart disease, and average glucose levels are important in stroke incidence. Notably, the research highlighted the increasing role of age, particularly in individuals over 45, reinforcing the need for targeted interventions in this demographic.

In evaluation of the performance of the four machine learning models—logistic regression, lasso regression, random forest, and XGBoost—it became apparent that traditional accuracy metrics could be misleading in the presence of class imbalance. For instance, although the random forest model attained the highest accuracy compared to others, its failure to predict any positive cases reveals a critical limitation. This discrepancy illustrates the necessity of implement F1 score and balanced accuracy as measurement of model performance, especially concerning the minority class.

The logistic regression model, despite the simplicity compare to other machine learning model, spectacularly exhibiting the strongest performance in balanced accuracy. This aligns with findings from Christodoulou, who reported no substantial evidence that advanced machine learning models achieve better outcomes than logistic regression, particularly in clinical prediction settings [7].

However logistic regression model still demonstrated challenges in identifying positive cases, suggesting that further refinement and optimization are required. The models' performances suggest that a multi-faceted approach, perhaps integrating different models or employing ensemble techniques, could enhance predictive accuracy and reliability.

Furthermore, this study reinforces the significance of early detection and risk factor management in stroke prevention. By utilizing machine learning models, healthcare providers can better identify at-risk individuals, facilitating timely interventions and potentially reducing the incidence of stroke.

5. Conclusion

In conclusion, this research underscores the critical role of machine learning models in stroke prediction, particularly through logistic algorithms. The findings reveal that while some models demonstrate high accuracy, they may fall short in correctly identifying positive cases, highlighting the importance of utilizing balanced metrics like F1 score and balanced accuracy in evaluating performance. The significant impact of risk factors such as hypertension, heart disease, and age on stroke incidence further emphasizes the need for ongoing research and model refinement.

The implications of this study offering practical insights for healthcare clinicians in improving stroke prevention strategies. By harnessing the power of machine learning, healthcare providers can better target interventions for high-risk populations, ultimately contributing to better health outcomes and reduced disability associated with stroke. Future research should improve the positive prediction accuracy by enhancing the models with more balanced data, exploring additional features, and investigating the integration of various predictive techniques to enhance overall performance and reliability in clinical applications.

References

[1] Centers for Disease Control and Prevention. (n.d.). 2024-10-1. Stroke facts. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/data-research/facts-stats/index.html.

[2] Goldstein, L. B., Adams, R., Alberts, M. J., Appel, L. J., Brass, L. M., Bushnell, C. D., Culebras, A., DeGraba, T. J., Gorelick, P. B., Guyton, J. R., Hart, R. G., Howard, G., Kelly-Hayes, M., Nixon, J. V., Sacco, R. L., American Heart Association, & American Stroke Association Stroke Council. Primary prevention of ischemic stroke: a guideline from the American Heart Association/American Stroke Association Stroke Council: cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group; Cardiovascular Nursing Council; Clinical Cardiology Council;

Nutrition, Physical Activity, and Metabolism Council; and the Quality of Care and Outcomes Research Interdisciplinary Working Group. Circulation, 2016, 113 (24), e873 – e923.

[3] Liu, Y., Yin, B., & Cong, Y. The probability of ischemic stroke prediction with a multi-neural-network model. Sensors (Basel, Switzerland), 2020, 20 (17), 4995.

[4] Ntaios, G., Faouzi, M., Ferrari, J., Lang, W., Vemmos, K., & Michel, P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. Neurology, 2012, 78 (24), 1916 – 1922.

[5] Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. Machine learning-based model for prediction of outcomes in acute stroke. Stroke, 2019, 50 (5), 1263 – 1265.

[6] Stroke Prediction Dataset. 2020. fedesoriano. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data.

[7] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 2019, 110, 12 – 22.