A Stroking Analysis of Strokes

Team 4: Andrew Kim, Benjamin Earnest, and Hunter Blum

Shiley-Marcos School of Engineering, University of San Diego

**Abstract**

Strokes are a common cause of death and disability across the world, costing billions in the United States every year (Murphy and Werring, 2020, p.1; Centers for Disease Control and Prevention 2022). Our project focused on constructing a sensitive and accurate model to identify what patients are most likely to have a stroke to capitalize on the success of early prevention techniques. The data was collected from Kaggle.com and consisted of 12 attributes describing traditional stroke risk factors (Palacios, 2021). After reading and cleaning the data in R, the cases of stroke were balanced using random duplicate oversampling (R Core Team, 2021). Each feature in the data was found to be important through the Boruta algorithm. 15 models were constructed using a variety of methods and mainly evaluated using accuracy and sensitivity. Based on sensitivity, a cost-sensitive CART with balanced data performed the best with a 97.6% sensitivity and 51.6%. However, an artificial neural network created with balanced data performed better overall, with a 74.4% accuracy and 90.2% sensitivity. Healthcare providers should note that the recommended artificial neural network has a low precision at only 13.0%. These models do not predict when a patient may have a stroke, they identify patients that are at a much higher risk of having a stroke. Therefore, this model should be utilized as a screener to find patients for further stroke testing, rather than a recommendation for prescribing medication or immediate, invasive stroke treatment measures.

# Predicting Stroke Risk Factors

# Caused by vascular injuries in the brain, strokes are the second leading cause of death and disability in the world (Murphy and Werring, 2020, p. 1). In the United States, somebody dies of a stroke every 3.5 minutes resulting in $53 billion in costs annually (Centers for Disease Control and Prevention, 2022). However, prevention strategies can have a massive impact, with current prevention techniques mitigating up to 80% of strokes (Pandian et al., 2018). A major part of beginning prevention is identifying patients who are at risk of stroke. Current model accuracies for predicting strokes in patients can range from 88% to 97% (Singh and Choudhary, 2017; Emon et al., 2020). This project considers existing data based on patient risk factors, identifies the key predictor variables, and evaluates model efficacy. The goal is to identify which modeling algorithm, provides the most accurate and sensitive predictions of strokes in patients when testing with novel data.

# Methodology

# Our goal is to develop a model to predict strokes in patients, utilizing a secondary data analysis. We utilized R to read in a data set from Kaggle.com, consisting of 5,110 records and 12 different attributes (R Core Team, 2021; Palacios, 2021). The arbitrary “id” attribute was removed from the data and attributes were converted to the correct type. 201 observations were missing BMI values and were subsequently removed from the data.

# To explore the data, we created a regular and normalized bar plot of each categorical attribute overlaid with the target stroke attribute. When normalized, the presence of heart disease or hypertension exhibited the largest relationship with having a stroke (Figure if we have room). For numeric attributes a regular and normalized histogram was coded. After normalizing, age had a clear, direct relationship with a stroke occurring in patients (Figure if we have room).

# The data was split into a training and testing dataset, making up 80% and 20% of the data respectively. The training data had a large imbalance for the target stroke attribute, with over 95% of observations not having a stroke. The stroke class was balanced in the training data using random duplicate oversampling, making each class of stroke equal. We selected important explanatory attributes by using the Boruta algorithm on the balanced data. Each variable had a much higher importance than the shadow data, so every attribute was considered in modelling (Figure 2.X).

**Figure 2.X.**

*Output of Boruta on balanced data. Explanatory attributes are in green and the shadow data are in blue.*

Chart, waterfall chart

Description automatically generated

The attributes considered for model development were gender, age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, BMI, and smoking status. A variety of models will be evaluated on a many metrics, particularly accuracy and sensitivity.

# C5.0

# The C5.0 model utilized all 11 variables, which generated multiple decision nodes. Figure TBD below features the C5.0 decision tree output from the Stroke data set. Based on the model, the “age” variable serve as the root node, which splits and determines if a patient is above or not above the age of a certain value. In this case, the first age split occurs at 44. The decision tree indicates that if a patient is less than or equal to the age of 44, it proceeds to the work type of the patient. If a patient is greater than the age of 44, the decision tree proceeds to another age value of 66, which determines the patients’ BMI and average glucose level. shows that the variable “age” serves as the root variable, which is split and determines the types of condition or employment people of certain age ranges have and whether it leads to having or not having a stroke. 66 nodes in total. Since the variables are packed tightly, a data frame was created that included the predictor variables of the records to classify. The C5.0 model was shown to generate a 78% accuracy with a sensitivity rate of approximately 54% and a specificity rate of approximately 80%.

**Figure TBD**

*C5.0 Decision Tree Predicting Stroke Causes.*

A picture containing chart

Description automatically generated

# CART

The CART model also utilized all 11 variables. However, the CART algorithm found the root node to be age, followed by avg\_glucose\_level, bmi, and work\_type as the decision nodes. This is similar to the C5.0 model. The CART model generated 9 leaf nodes in total. The root node considered age greater than 45 as an initial decision point, with only 28% being below age 45, and low risk of having a stroke. The remaining 72% were then split for those above or below age 67. The model considered anyone over the age of 67 to be at risk of a stroke, suggesting age is a leading factor in determine stroke risk. There were 35% of this group between the ages of 45 and 67, that would need other factors considered to determine stroke risk by the CART model. For this age group, the model considered avg\_glucose\_level, bmi, and if a person was Self-employed. Interestingly, the model found that a person who was between ages 45 and 67, had a bmi greater than 32, and is Self-employed represented a stroke risk. The plot for the CART model is in Figure TBD shows the CART decision tree plot. When evaluating the model, we found it to have an accuracy of 80%, sensitivity of 89%, and specificity of 70%. This suggests the model generalized well, and has the potential for future predictive model use.

**Figure 2.X.**

*Output of CART model.*



# Logistic Regression

Given the binary nature of our target variable and many of the predictor variables, a logistic regression model will be considered and performance evaluated. To do this, the non-binary continuous variables will be standardized. This included age, bmi, and avg\_glucose\_level. Once standardized, and initial run of the logistic regression model was run to look for statistically significant predictor variables. To avoid issues from an imbalanced data set, the balanced, standardized training data set was used. The model found that gender, hypertension, heart\_disease, smoking\_status, age, and avg\_glucose\_level were statistically significant, or had p-values < 0.05. The algorithm was then rerun to validate statistical significance of the variables and generate the logistic regression equation for our model. Equation 1 shows the model developed.

The test data was run through the model to evaluate performance. The results suggest the model does not predict well based on the variables used, as each test observation was predicted to have a stroke, or greater than a 50% chance of having a stroke.

# Random Forest

Two random forest models were created utilizing the “rf” method in the caret package (Kuhn, 2021). The first model utilized the unbalanced data, while taking into account the class probabilities of the stroke. The final model had an accuracy of 95.8% on the test data. However, the unbalanced model had a sensitivity of zero. In hopes of improving the sensitivity, another random forest model was produced using the balanced data set. However, this model resulted in the same output, resulting in an accuracy of 95.8% and a sensitivity of zero.

# Naïve Bayes

# Several Naïve Bayes models were created and evaluated utilizing different predictor variables. When assessing the Naïve Bayes model, the first model that evaluated “stroke” in association with “gender” and “hypertension” generated the highest accuracy rating of 88% and the highest specificity rating of 91%. Based on this, this indicates that evaluating the gender type and hypertension level of each patient are leading indicators of a patient to have a stroke. The second model that evaluated “stroke” in association with “heart disease” and “ever married” generated the highest sensitivity rating of approximately 88%.

# Neural Network

Four neural network models were created using the “nnet” method in the caret package and ten-fold cross-validation. Two models were made based on the unbalanced data, one with the raw data and the other with z-score standardized numeric attributes. Both models resulted in all-negative models, producing 95.8% accuracy and zero sensitivity when used with the test set. The third model utilized balanced data. When run on the test data set, the balanced neural network had an accuracy of 78.7% and an 80.5% sensitivity. Finally, a fourth model was made, this time using the balanced data with z-score standardized numeric attributes. This model had a lower accuracy at just 61.3%, but a higher sensitivity at 87.8%.

# Association Rule

# Results

# The models were evaluated with accuracy, sensitivity, specificity, precision, and F1 score against a baseline all-negative model. The unbalanced artificial neural network (ANN Reg.), unbalanced and standardized artificial neural network (ANN Z), and unbalanced random forest (RF Reg.) all produced all negative predictions, equaling the Baseline model in every measure (Table X.X). The balanced random forest (RF Bal.) model generated the second-highest accuracy rate of approximately 95%, however, the model also did not predict any patients that actually had a stroke correctly. Together, all models generated the highest proportion of predicted positives to actual positives. The strongest model according to sensitivity was the cost-sensitive CART (CART Cost Bal.), which generated the highest sensitivity at 97.6%. The highest precision rating was found through Logistic Regression at approximately 75%, which indicated the highest proportion of predicted strokes to actual strokes. The balanced artificial neural had the highest F1 score at 23%.

**\*Note: All were made through the final model evaluations and when evaluating the comparison table.**

**Table 1**

*Model Evaluation Table for Stroke Prediction*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **F1** |
| **Baseline** | 0.958 | 0.000 | 1.000 | 0.000 | 0.000 |
| **Log Reg.** | 0.042 | 1.000 | 0.000 | 0.418 | 0.080 |
| **ANN Reg.** | 0.958 | 0.000 | 1.000 | 0.000 | 0.000 |
| **ANN Bal.** | 0.744 | 0.902 | 0.737 | 0.130 | 0.228 |
| **ANN Z** | 0.958 | 0.000 | 1.000 | 0.000 | 0.000 |
| **ANN Z Bal.** | 0.613 | 0.878 | 0.602 | 0.088 | 0.160 |
| **RF Reg.** | 0.958 | 0.000 | 1.000 | 0.000 | 0.000 |
| **RF Bal.** | 0.948 | 0.000 | 0.989 | 0.000 | 0.000 |
| **CART Bal.** | 0.700 | 0.902 | 0.691 | 0.113 | 0.201 |
| **CART Cost** | 0.845 | 0.512 | 0.860 | 0.138 | 0.218 |
| **CART Cost Bal.** | 0.516 | 0.976 | 0.496 | 0.078 | 0.144 |
| **C5.0 Bal.** | 0.785 | 0.537 | 0.796 | 0.103 | 0.173 |
| **NB Gender** | 0.885 | 0.268 | 0.912 | 0.117 | 0.163 |
| **NB Heart + Marry** | 0.376 | 0.878 | 0.354 | 0.056 | 0.105 |
| **NB Resident** | 0.505 | 0.512 | 0.505 | 0.0432 | 0.080 |
| **NB Smoke + Work** | 0.330 | 0.854 | 0.308 | 0.051 | 0.096 |

# Chart, bar chart Description automatically generated

Figure X.X. Accuracy of each model, with grey meeting the baseline and red performing worse.

Chart

Description automatically generated

Figure X.X. Sensitivity of each model, with grey meeting the baseline and green performing better.

# Conclusion

Our goal was to find a model with high accuracy and sensitivity to predict if a patient would have a stroke. There are a few models that we would recommend based on the circumstances. If we were solely focused on capturing the most people who would actually have a stroke, we would recommend the balanced, cost-sensitive CART (CART Cost Bal.) This model had the highest sensitivity, however, it also struggled to correctly predict patients who would not have a stroke, resulting in a poor accuracy (Figure X.x). If we were focused on a more balanced model, the balanced artificial neural network (ANN Bal.) would be our recommendation. This model still had a high sensitivity at 90.2%, and had a much higher accuracy than the CART Cost Bal. at 74.4%. Choosing one, we would recommend the ANN Bal. because of its better accuracy. However, it that the model has a low precision at just 13%, meaning that 87% of patients predicted as having a stroke by this model did not have a stroke. Healthcare providers should take this into account if using this model to suggest any medications with potentially harmful side effects. The model does serve as a good screener to capture all patients that will have a stroke, so healthcare providers can use this to narrow down patients for more advanced stroke risk evaluation.

# References

Centers for Disease Control and Prevention. (2022, April 5). Stroke facts. Centers for Disease Control and Prevention. Retrieved April 13, 2022, from https://www.cdc.gov/stroke/facts.htm

Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., & Kaiser, M. S. (2020, November). Performance analysis of machine learning approaches in stroke prediction. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1464-1469). IEEE.

Markus, H. (2008). Stroke: causes and clinical features. Medicine. 36(11). Elsevier Ltd.

Murphy, S. J., & Werring, D. J. (2020). Stroke: causes and clinical features. Medicine, 48(9), 561-566.

Palacios, F. S. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved April 14, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Pandian, J. D., Gall, S. L., Kate, M. P., Silva, G. S., Akinyemi, R. O., Ovbiagele, B. I., ... & Thrift, A. G. (2018). Prevention of stroke: a global perspective. The Lancet, 392(10154), 1269-1278.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Singh, M. S., & Choudhary, P. (2017, August). Stroke prediction using artificial intelligence. In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON) (pp. 158-161). IEEE.