A predictive analytics for stroke prediction using data mining approaches

Team 3: Reina Chehayeb, Yuqi Chen, Safwaan Karim, John Marko, Jiale Wang
Fordham University, Gabelli School of Business
Data Mining for Business, ISGB-7967-001 Fall 2022, Dr. Lin Hao

Abstract

A stroke, also known as transient ischemic attack or cerebrovascular accident, happens when blood flow to the brain is blocked. According to World Health Organization (WHO), stroke is the second leading cause of death and the third leading cause of disability globally. Every year, more than 795,000 people in the United States have a stroke. Every 3.5 minutes, someone dies of stroke.[1] If we can anticipate stroke before it occurs and can take effective preventive measures for individuals at risk, it is highly possible that the incidence of stroke can be reduced. In addition, predictive analytics of stroke would assist practitioners in determining the risk of experiencing a stroke on a per-patient basis, and optimize screening and preventative care protocols. Therefore, from a business perspective, stroke prediction would assist the operations of hospitals, clinics, insurance companies, and other medical organizations by minimizing unnecessary expenses and use of resources. This project analyses the various factors of a health records dataset downloaded from Kaggle.com[2]. Using various data mining techniques, we identify the most important factors for detecting a stroke (i.e. age, average glucose level, and hypertension). Furthermore, by comparing different model approaches, we conclude that Logistic Regression with Cross Validation and SMOTE, as well as Nearest neighbor are the two best models for stroke prediction. Each model has its own strengths and shortcomings for business use cases, which are further elaborated upon in this paper.

[1] https://www.cdc.gov/stroke/facts.htm

[2] https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5/data

Introduction

I. Data Acquisition and Dataset Description

In this project, we retrieved a dataset from Kaggle, entitled "*Stroke Prediction Dataset*". It was originally published to the website on May 5th, 2021 by a user named Josh.

This dataset contains the health records of 5,110 patients with 12 attributes. There are 10 independent variables, 1 unique identifier for each patient, and 1 dependent variable. The dependent variable is binary, and indicates the occurrence of a stroke; a '0' indicates that a stroke did not occur, while a '1' indicates that a stroke occurred. The independent variables are: gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, and smoking_status. A detailed description of each attribute is shown in Table 1.

Variable name	Variable description	Variable type
stroke*	Occurence of stroke*	Binary (0, 1)*
gender	Gender	Binary/Categorical (Male, Female, Other)
age	Age, in years	Numerical
hypertension	Presence of hypertension as pre-existing condition	Binary (0, 1)
heart_disease	Presence of heart disease as pre-existing condition	Binary (0, 1)
ever_married	Marital history/whether or not the individual is married	Binary (0, 1)
work_type	Nature of the individual's work	Categorical (Children, Government job, Private, Never worked, Self employed)
residence_type	Urbanacity of residence	Categorical (Urban, Rural)
avg_glucose_level	Average glucose level in blood	Numerical
bmi	Body Mass Index	Numerical
smoking_status	Whether or not the individual smokes	Categorical (Formerly smoked, Never smoked, Smokes, Unknown)

Table 1: Attributes Description (* denotes the dependent variable)

The rationale for selecting this specific dataset is built upon its sound design and its comprehensive content. Firstly, the dataset consists of historical observations organized as individual level data, where each column contains one variable and each row is an observation for an individual, as opposed to aggregated data. Additionally, the dependent variable has observations that show the cases where the events of our interest did happen. Further, all the independent variables are relevant to the occurrence of a stroke by using institutional knowledge. Further relevance detections will be shown in Section II Data Exploration. Finally, the number of observations in the dataset satisfies the portrait-shape requirement. The minimum number of observations is:

$$n = 6 \cdot m \cdot (u + 1) = 6 \cdot 2 \cdot (10 + 1) = 132$$

Here, m indicates the number of classes in the dependent variable stroke and u indicates the number of predictors. The dataset contains 5,110 observations, which greatly exceeds the minimum requirement. It is important to note that, as the source of the data is confidential, we are not aware of the exact data collection methods employed in constructing the dataset. Therefore, we assume all the predictors to be exante.

II. Data Exploration[3].

For numerical predictors, we generated side-by-side box plots to display any relevant relationships with the dependent variable stroke, as shown in Figure 1.

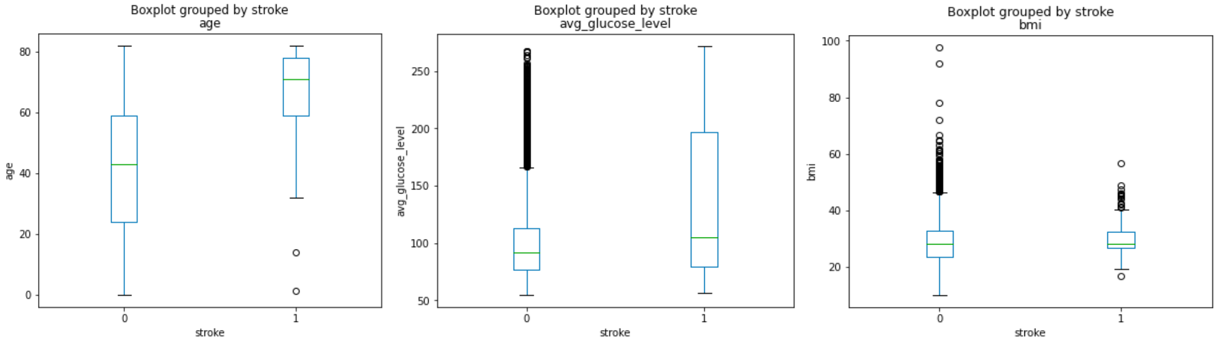

Figure 1: Side-by-side Boxplots of Stroke vs. Age, Average Glucose Level and BMI

For categorical predictors, we conducted Chi-square test of independence. The null hypothesis H_0 set to indicate no relationship between two categorical variables. The alternative hypothesis H_1

[3] Data used in Data Exploration Section is the cleaned data. Details of the data cleaning process refers to Method – Data Preparation Section.

supported a relationship between two categorical variables. The Chi-square test was conducted at the $p \leq 0.05$ significance level. If the results met this significance requirement, the null hypothesis can be rejected. Otherwise, we would fail to reject the null hypothesis, and there may be interactions between predictors.

With ‘gender’, ‘hypertension’, ‘heart_disease’, ‘ever_married’, and ‘residence_type’ all having 2×2 cross tables, we conducted the Chi-square test using scipy.stats.chi2_contingency, as shown in Table 2.

	χ^2-value	p-value	Expected Frequencies
gender	0.340002536046177	0.5598277580669416	[[2848.07985907, 145.92014093], [2011.92014093, 103.07985907]]
hypertension	81.57314462043591	1.688936253410575e-19	[[4386.27128597, 224.72871403], [473.72871403, 24.27128597]]
heart_disease	90.22943664078221	2.120831133146208e-21	[[4597.45155608, 235.54844392], [262.54844392, 13.45155608]]
ever_married	58.86780922483486	1.6862856191673454e-14	[[1670.41691133, 85.58308867], [3189.58308867, 163.41691133]]
Residence_type	1.0749713079092142	0.29982523877153633	[[2390.52260716, 122.47739284], [2469.47739284, 126.52260716]]

Table 2: Chi-square Tests for Gender, Hypertension, Heart Disease, Maritual Status, Residence Type vs. Stroke

Since all of the expected frequencies are greater than 5, the chi-sqaure test results are considered credible. For ‘hypertension’, ‘heart_disease’, and ‘ever_married’, we can reject the null hypothesis as the p-value is less than 0.05. Thus, the result indicates that there is a relationship between stroke and hypertension, heart disease, or marital status. For ‘gender’ and ‘residence_type’, we fail to reject the null hypothesis as the p-value is greater than 0.05. Thus, the result indicates that there is no or weak relationships between stroke and gender or residence type.

With ‘work_type’ and ‘smoking_status’ both having chi-square tables larger than 2×2, we conducted a post hoc testing using Bonferroni-adjusted-p-value method. The adjusted p-value for ‘work_type’ is $0.05/5 = 0.01$ and the adjusted p-value for ‘smoking_status’ is $0.05/3 \approx 0.017.$, as shown in Table 3 and Table 4.

work_type	**χ^2-value**	**p-value**
Govt_job	0.008660334140911895	0.9258551669275809
Never_worked	0.32242006724269884	0.570156890083992
Private	0.6191734473868207	0.4313546039669218
Self-employed	18.955485547287754	1.3380401602010203e-05
children	34.82046070158722	3.615537224238463e-09

Table 3: Multiple Chi-square Tests for Work Type vs. Stroke

smoking_status	**χ^2-value**	**p-value**
formerly smoked	20.588666991495394	5.6932145282837105e-06
never smoked	17.20949941122615	3.3475824871454873e-05
smokes	0.2999841198085903	0.5838923763940107

Table 4: Multiple Chi-square Tests for Smoking Status vs. Stroke

In Table 3, for ‘self-employed’ and ‘children’ being a work type, we can reject the null hypothesis as the p-value is less than 0.01. Thus, the result indicates that there is a relationship between stroke and self-employed or children. For ‘govt_job’, ‘never_worked’, and ‘private’ being a work type, we fail to reject the null hypothesis as the p-value is greater than 0.01. Thus, the result indicates that there is no or weak relationship between stroke and having a government job, never worked, or having a private work.

In Table 4, for ‘formerly smoked’ and ‘never smoked’, we can reject the null hypothesis as the p-value is less than 0.017. Thus, the result indicates that there is a relationship between stroke and formerly smoked or never smoked. For ‘smokes’ being the smoking status, we fail to reject the null hypothesis as the p-value is greater than 0.017. Thus, the result indicates that there is no or weak relationship between stroke and smokes.

At this point of data exploration, we conclude that there is a strong relationship between age and the occurrence of stroke, whereas gender, residence type, some sub-classes of work type and smoking status may not be strongly relevant to the occurrence of stroke. However, due to institutional knowledge, we decide to include all of the predictors into consideration for our further data mining approaches. Several conclusions made here will be further manifested by the logistic regression with pre-specified alpha approach in the following sections.

III. DIDA Framework

Data.

The data we use for this project is obtained from a Kaggle database, as described in Section I. The information provided includes demographic and behavioural variables, the presence of pre-existing health conditions, and whether or not a stroke occurred.

Insights.

In analyzing the data, we aim to determine the probability of a stroke occurrence (indicated as a binary event).

Decision.

The insight will inform screening protocols for hospitals, clinics, and other medical organizations. Practitioners could use the model mentioned later in this paper to determine the risk of experiencing a stroke on a per-patient basis, and decide whether or not to screen those patients or implement preventative care. The model could also help determine insurance costs. Insurance companies could use the risk level of individuals to calculate their insurance costs.

Advantage.

Hospitals and clinics would minimize expenses of screening and treatment by narrowing down the pool of patients for which these procedures should be administered.

Methods

Our project comprised 5 phases, shown in Figure 2 below.

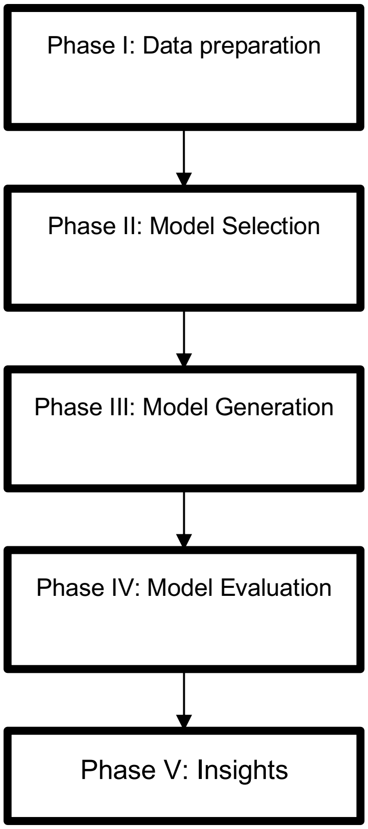
Phase I: Data preparation
Phase II: Model Selection
Phase III: Model Generation
Phase IV: Model Evaluation
Phase V: Insights

Figure 2. 5-Phase plan for analysis

I. Data preparation.

To prepare the dataset for analysis, we first carried out an evaluation of each attribute, and then standardized the data.

First, we removed the identifying attribute ('ID') as it did not offer valuable information for the purpose of our analysis. We then examined the remaining data for null values. 201 missing values were found for the BMI predictor. To address this, we imputed the median BMI value to replace any missing values. It is of interest to note that removing these records altogether would have still yielded a sufficiently large dataset. However, the proportion of records with missing BMI values that experienced a stroke (i.e., the outcome of interest) was higher than that of the overall dataset. Therefore, it was determined that the removal of these records could have impacted the results of the analysis significantly. Whether or not the presence of data for one’s BMI and the occurrence of a stroke are correlated (which is potentially the case given the difference in the proportion of stroke occurrence between those who had BMI data and those who did not) is outside the scope of this study. However, knowledge of and/or willingness to disclose health information may be a predictor of interest in future analysis.

Next, we searched the dataset for irregular or unexpected values by generating all the unique values that these attributes took on in the dataset. We found that one record had a value of ‘Other’ for the Gender attribute. All other records had binary values for this predictor. For simplicity, this record was removed from the dataset. Due to this being the only record with this value, no valuable conclusions could have been drawn for members of this group based on one observation. It is important to note that further research into the likelihood of experiencing a stroke for marginalized populations is needed.

In addition to the Gender attribute, we explored the Smoking Status attribute further, since 1,544 records had a value of ‘Unknown’ smoking status. We replaced all ‘Unknown’ values with ‘Never Smoked’, the mode value for this predictor.

Finally, all numerical values were scaled to their respective standardized values, and all categorical values were standardized via dummy coding procedures included in the scikit-lean library in Python.

II. Model selection.

Given the nature of our dataset, our outcome of interest qualifies as a rare event; the proportion of records that represent the occurrence of a stroke being much lower than that of records that represent no stroke occurrence. Only 4.8% of the records in the dataset indicate the occurrence of a stroke. This causes issues in training a model, since the outcomes represented in the dataset are so greatly skewed.

To combat this, we applied the Synthetic Minority Oversampling Technique (SMOTE). SMOTE works by creating additional data points for the underrepresented outcome. These data points are created by considering the nearest neighbors to the data points that represent the outcome of interest. The vector between the existing point and a nearest neighbor is multiplied by a value between 0 and 1. This is added to the existing data point to create a new, synthetic data point.

We developed three models, using logistic regression with cross-validation (LRCV), Nearest Neighbors approach (kNN), and classification trees (CT). LRCV and CT were developed twice: once before applying SMOTE, and once after. Due to the nature of the kNN model, SMOTE was not applied.

The volume of the sample size, the speed of training the model, and the handling of outliers were considered in selecting these approaches. While a Neural Network model was considered, the dataset may not have been large enough to produce a useful model. For this reason, and because Neural Networks offer low interpretability, we chose not to generate a Neural Network model. Additionally, since the missing data was handled appropriately, there were no concerns around the sensitivity to missing data that is common in the LRCV and kNN approaches.

Further, all three models would allow us to predict whether or not an individual is likely to experience a stroke. The LRCV and CT provide additional insight into the predictors that contribute most to that likelihood. The kNN model does not provide this insight (i.e., kNN has a lower outcome interpretability), but can yield accurate predictions, and so was included in the analysis.

III. Model Generation.

Using the train_test_split function from the scikit-learn library, the data was randomly partitioned into two sets: a training set containing 80% of the data, and a testing set containing the remaining 20% of the data. The models from Phase II were trained on the training set after applying SMOTE. Then, the models were tested against the testing set to determine the accuracy of each model. There was no indication to apply SMOTE to the testing set. All models were trained using functions from the scikit-learn library in Python.

LRCV.

Redundant dummy variables were dropped from the dataset for the logistic regression models. The logistic regression model was first run without cross-validation, at alpha = 0.8. The results of this were used to determine the importance of the predictors before running the proper LRCV. For the

LRCV, the model was trained across 5 k-folds of the training set, with alphas ranging from 0.001 to 100. All processors were used to train the model (i.e. n_jobs argument was set to -1) to reduce the time required to train the model. Additionally, the solver was set to ‘saga’ to achieve better convergence of the model. This model converged with less than 5,000 iterations.

The LRCV model was trained twice, once before applying SMOTE and once after. All parameters, including the k-folds, number of processors used, solver used, and iterations were controlled to remain the same to isolate the effect of applying SMOTE.

kNN.

Redundant dummy variables were not dropped from the dataset for this model. The kNN model was initially trained without using cross-validation to determine its initial accuracy. For this, the k nearest neighbors was set to 5, and utilizing Euclidean metrics to account for the geometric distances between data points. Then, like the LRCV model, the kNN model was trained to find the optimal value for k using cross-validation, with the maximum k set to 200.

CT.

Redundant dummy variables were dropped from the dataset prior to generating the CT. The tree model was generated twice– once before applying SMOTE, and once after. The CT was generated across 5 k-folds, and based on a range of depths from 1 to 100. These parameters remained consistent for both CT models.

IV. Model Evaluation.

The LRCV models were evaluated based on their accuracy and confusion matrices (i.e., proportion of correctly identified outcomes and incorrectly identified outcomes). The kNN and CT models were evaluated based on their accuracy. Accuracy was measured against the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC). The AUC was calculated using the roc_auc function of the scikit-learn library.

V. Insights.

In addition to comparing the accuracy of each model to determine the most accurate one, the characteristics of each model were analysed to determine which would be most useful given the medical and severe nature of the outcome. For example, if a model had high accuracy as a result of predicting no stroke across the board (i.e. the model generated was a naive model), it would not be as useful as a model that predicted both potential outcomes with slightly lower accuracy. Similarly, if a model erred on the side of caution and slightly over-predicted the occurrence of a stroke, that characteristic would be preferred.

The results from each model were also analysed to determine relationships between predictors and the outcome, where possible (i.e., in the LRCV and CT models, but not the kNN).

Results

LRCV.

Before applying SMOTE, the LRCV model showed an optimal alpha level of $\alpha = 2.103$. At this level, the variables and their corresponding beta coefficients were as outlined in Table 5. BMI, as well as some predictors related to work type and smoking status, were dropped from the model at this alpha level.

The most important predictor in the LRCV model without SMOTE applied was age, which remained in the model up to a penalty level of $\alpha = 90$, followed by glucose levels, which remained in the model up until $\alpha = 15$. The third most important predictor was hypertension, which remained in the model up until a penalty level of $\alpha = 13$. The least important predictor was a work status indicating never having worked, followed by being a (current) smoker, and BMI. These three predictors dropped from the model at a penalty level of $0.05 < \alpha < 0.1$.

Predictor	**Beta Coefficient**
Age	1.596337
Hypertension_Yes	0.331337
Heart_Disease_Yes	0.229323
Avg_Glucose_Level	0.183235
Work_Type_Private	0.194559
Residence_Type_Urban	0.104209

Gender_Female	0.066077
Smoking_Status_Never Smoked	-0.078182
Work_Type_Self Employed	-0.096496
Ever_Married_Yes	-0.385241
Smoking_Status_Smokes	0.000000
Work_Type_Children	0.000000
Work_Type_Never Worked	0.000000
BMI	0.000000
Intercept	-3.909809

Table 5. Beta coefficients for LRCV without SMOTE

The confusion matrix for the LRCV model without SMOTE applied is shown in Table 6, indicating an accuracy of 94.129%. Here, a 1 indicates the occurrence of a stroke, whereas a 0 indicates no stroke.

	Predicted 0	**Predicted 1**
Actual 0	962	0
Actual 1	60	0

Table 6. Confusion matrix for LRCV without SMOTE

After applying SMOTE, the LRCV model showed an optimal alpha level of $\alpha = 0.602$. At this level, the variables and their corresponding beta coefficients were as outlined in Table 7. The most important predictor was Age, consistent with the previous model.

Predictor	**Beta Coefficient**
Age	2.051125
Hypertension_Yes	-0.334817
Heart_Disease_Yes	-0.579004
Avg_Glucose_Level	0.186028

Work_Type_Private	-0.208657
Residence_Type_Urban	-0.184123
Gender_Female	-0.118626
Smoking_Status_Never Smoked	-0.846583
Work_Type_Self Employed	-0.847223
Ever_Married_Yes	-0.753097
Smoking_Status_Smokes	-0.950712
Work_Type_Children	-0.306506
Work_Type_Never Worked	-0.362236
BMI	-0.009840
Intercept	0.441795

Table 7. Beta coefficients for LRCV with SMOTE

The confusion matrix for the LRCV model with SMOTE applied is shown in Table 8, indicating an accuracy of 74.417%. With SMOTE, the LRCV model predicted significantly more occurrences of stroke, but yielded a lower accuracy. The implications of this are further analysed in the Discussion.

	Predicted 0	**Predicted 1**
Actual 0	715	247
Actual 1	17	43

Table 8. Confusion matrix for LRCV with SMOTE

kNN.

For the nearest neighbors model, the optimal k (i.e., number of nearest neighbor data points used to make a prediction) was 185. The accuracy of this model over the test partition was 83.570%, considerably higher than the LRCV with SMOTE, but slightly lower than the LRCV without SMOTE.

CT.

Before applying SMOTE, the CT model generated the best pruned tree with a depth of 2 and accuracy of 83.669%. At this level, the tree is shown in Figure 3.

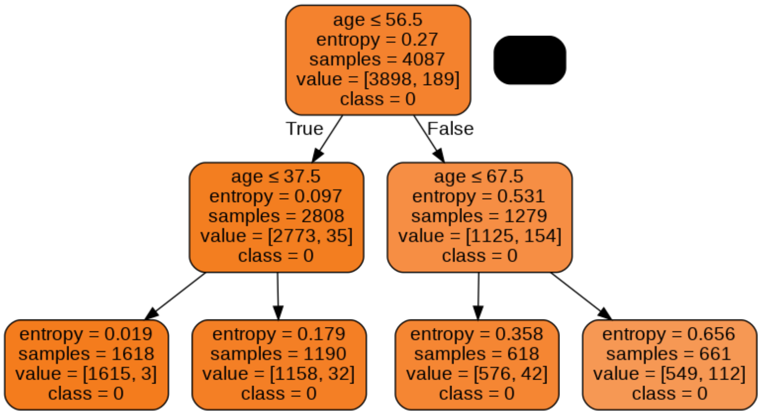

Figure 3: Best Pruned Tree before SMOTE

The main differentiating predictor in the above tree is ‘age’. Since our dataset is highly skewed in its outcomes (i.e. susceptible to a rare-event problem), only 5% of observations have ‘stroke’ valued as ‘1’, and the remaining 95% of observations valued as ‘0’s. The best pruned tree was generated as a naive classification model, predicting all ‘0’s at the terminal nodes in order to maximize its accuracy. While this model is accurate, it is not useful. Therefore, to rectify this problem, we applied SMOTE to the CT model.

After applying SMOTE, the CT model generated the best pruned tree with a depth of 2 and accuracy of 82.554% (slightly less accurate than without SMOTE). The final selected tree is shown in Figure 4.

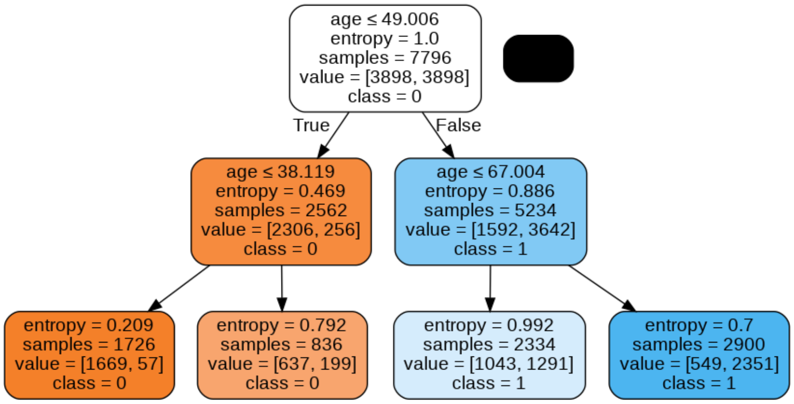

Figure 4: Best Pruned Tree after SMOTE

The differentiating predictor in the above tree is also ‘age’. However, with SMOTE, the tree is more useful since it predicts both ‘0’s and ‘1’s at the terminal nodes. Four rules were generated in total. We indexed them as leaf node 1 to 4 (the leftmost leaf node to the rightmost leaf node). When evaluating these rules, we firstly ranked their effectiveness by calculating the rules’ predicted probability of the event of interest, as shown in Table 9. Secondly, we ranked their significance by comparing the number of training observations that support the rule, as shown in Table 10.

Leaf Node #	**Effectiveness**	**Rank**
Leaf Node 1	0.97	1
Leaf Node 2	0.76	3
Leaf Node 3	0.55	4
Leaf Node 4	0.81	2

Table 9: Rules’ Effectiveness Ranking

Leaf Node #	**Significance**	**Rank**
Leaf Node 1	1669	2
Leaf Node 2	637	4
Leaf Node 3	1291	3
Leaf Node 4	2351	1

Table 10: Rules’ Significance Ranking

Lastly, by applying the rule of thumb for rule reporting, we picked the top 2 effective English rules (Node 1 and Node 4) and selected the one with the highest significance (Node 4) as our final English rule: $IF\ age > 67, THEN\ Predict\ stroke = 1$. For business purposes, that prediction may indicate a screening or insurance pricing premium.

Discussion

The LRCV model that was generated before applying SMOTE had the highest accuracy, followed by the CT, kNN, and finally, the LRCV after applying SMOTE. However, in addition to considerations around accuracy, it’s important to determine the functionality of each model as it aligns with our DIDA framework, and the extent to which each model supports the final advantage. Holistically, the best model generated in terms of accuracy was the kNN model, and the best model generated to inform next steps for patients was the LRCV model generated after applying SMOTE.

While the LRCV model without SMOTE applied had, by far, the highest accuracy, it behaved as a naive model; it always predicted no stroke. Thus, we concluded that the accuracy was not valid for the purpose of our analysis. It would reduce the number of unnecessary screenings, which is technically aligned with our DIDA framework, but if the ultimate goal of the hospital is to successfully screen patients, choosing not to screen anyone is not practical. Given the purpose of our analysis, it is better to be overly cautious and sacrifice some accuracy for a more useful model. Thus, the LRCV model generated before applying SMOTE was not considered to be the strongest model from our analysis.

Similarly, the CT models showed high accuracy, but only took age into account as a predictive factor. This model may not be very helpful, but could provide a simple guideline for patients above a certain age (ex. to conduct additional stroke screening with other routine assessments if the patient meets the age threshold).

Following these considerations, the LRCV model (with SMOTE applied) and the kNN model remain. The LRCV with SMOTE, while lower in accuracy (approximately 74%), provides a good understanding of the relationship between the predictors and the outcome of interest. Alternatively, the kNN model offers a significantly higher accuracy (approximately 83%), but due to the characteristics of the kNN approach, does not offer insight into the relationship between the predictors and the outcome.

To determine the best model, the DIDA framework must be revisited to qualify the benefits of understanding the predictor-outcome relationship. The models were developed to drive the decision-making process around screening patients for stroke or pricing insurance, and the advantage offered by utilizing the model is narrowing down the pool of patients to screen. So, while the kNN model is more beneficial in determining which patients to screen, the LRCV model is useful in narrowing down the most important factors to address to reduce the risk of stroke.

The analysis contains some limitations that must be addressed. Firstly, only one specific library in Python was used to generate each model. While scikit-learn is a reputable and very widely used library, it may be helpful to compare the same type of model as it would be generated by functions from different libraries, or even in different programming languages. Additionally, there were limitations around the predictors included in the dataset. Some predictors that were not included in the dataset, such as cholesterol levels, may be beneficial to capture in future analyses.