Stroke Prediction

Hasani Pathirana, Mary Solomon and Corey Thrush

STAT 6440: Data Mining

Dr. Shuchismita Sarkar

April 24, 2021

1. Introduction

1.1 Motivation

With the advancement of technology, people have become busier and less attentive to their heath. Because of this, they have created a very stressful environment and are suffering from lots of heart diseases. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Stroke is typically understood to be at high risk for older adults, however medical experts are citing the rise of stroke in younger patients. Hence identifying the risk factors for causing stroke has become one of the leading topics that needs to provide careful attention.

1.2 Objectives

Our intent of analyzing this data set is to accurately classify those who had a stroke and identify the high risk factors that contribute to this outcome. Thus, identifying those risk factors can contribute towards creation of a public service announcement by promoting healthy activities. Furthermore, the identification of these risk factors can allow for early detection and prevention of experiencing a stroke. Additionally, we want to compare many models' performance.

1.3 Data

The dataset used for this project is the “Stroke Prediction Dataset” from Kaggle where each of the 5110 rows represents health and lifestyle observations for a patient. The dataset has twelve variables overall, but when excluding the id variable, there are ten predictors and one response. A description of the variables in the dataset are in the table below:

Variable name	**Description**
id	unique identifier
gender	"Male", "Female" or "Other"

age	age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	average glucose level in blood
bmi	body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
stroke	1 if the patient had a stroke or 0 if not

Again, we consider stroke as our response variable and we have 249 individuals who had a stroke and 4861 of them not having a stroke.

2. Methodology

In order to explore the best method for predicting stroke in patients, a variety of classification methods including logistic regression and random forest will be compared. Furthermore, methods such as variable selection with stepwise regression and regularized regression with ridge, lasso and elastic net will be modeled. These classification approaches will show how the mix of numeric and categorical predictors contribute to classifying the patients as stroke or non-stroke. In addition, the nature of the stroke classifier creates a situation of a rare event as there is a much greater proportion of patients to not suffer from a stroke. To treat the rare event nature of stroke classification comparisons of model performance between undersampling and oversampling methods will be explored.

Logistic regression, or binary logistic regression, models multiple numeric and categorical predictors to classify a binary response outcome to two possible outcomes. In the case of this project, we will be modeling the mix of health variables in classifying the patient as being prone to a stroke or not. The logistic regression model is expressed as :

$$log(\frac{\pi(y)}{1-\pi(y)}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Where π is the probability that the response variable indicates that the patient will experience a stroke. The probability π takes values from 0 to 1 and is modeled as the logit transformation, $log(\frac{\pi}{1-\pi})$, to form a linear function of the predictors. The logistic regression model can be optimized by methods of variable selection and regularized regression.

Variable selection aims to reduce the number of predictors, while maintaining the performance and increasing interpretability. The variable selection used is the stepwise backwards elimination method. This method initially fits the full model, then removes the predictor with the smallest contribution if it lacks statistical significance. This process is repeated until all regressors or if the contribution of the model is statistically significant.

An alternative method of optimizing the logistic regression method is to fit the data to regularized regression. This family of methods aims to shrink the estimates towards zero by minimizing the SSE and adding a penalty factor in order to reduce the variance of the model. Ridge regression implements this penalty by minimizing SSE and adding the penalty term of

$\lambda \sum_{j=1}^{k} B_j^2$. Ridge will shrink coefficients towards zero, but never exactly zero. Therefore, the Lasso method is a regularized regression technique that also performs variable selection. It minimizes SSE, but adds a penalty in the form of $SSE + \lambda \sum_{j=1}^{k} \left|B_j\right|$. Unlike Ridge, Lasso may shrink and force coefficients to approach zero, therefore performing variable selection. The final regularized regression method is Elastic Net, which combines features of Ridge and Lasso by shrinking the coefficients of correlated variables together while also performing variable selection. The penalty term added to the minimized SSE is $\lambda \sum_{j=1}^{k} (\alpha\left|B_j\right| + (1 - \alpha)B_j^2)$.

Finally, the method to compare with the various logistic regression approaches is Random Forest. Random Forest differs from the regression approach such that it is an ensemble method which creates predictions on an average of the predictors obtained from B bootstrap samples. For classification, this means the majority class label among $T_1(x), T_2(x), \dots, T_B(x)$ trees is the final predicted classification for an observation. Furthermore, for each split of the bootstrapped classification trees, not every predictor is considered. Rather, the optimal number of predictors to consider at each each split is $\sqrt{p}$, where p is the number of predictors.

To properly fit and evaluate the models, training and test data are used. Such that training data is fit to the model where the fitting occurs with 10 fold cross validation for all of the logistic regression models. Then, the performance of these models are evaluated on the test data. The performance of all classification models will be evaluated using the measures of accuracy, kappa, sensitivity and specificity. Considering the following confusion matrix:

	Actual C1	Actual C2
Predicted C1	True Positives (TP)	False Positives (FP)
Predicted C2	False Negatives (FN)	True Negatives (TN)

The performance measures are defined:

- Accuracy: The proportion of predicted values that match their true class. $\frac{TP + TN}{TP + FP + FN + TN}$
- Kappa: Metric that compares an Observed Accuracy ($p_o = \frac{TP + TN}{TP + FP + FN + TN}$) with an Expected Accuracy ($p_e = p_{c_1} + p_{c_2}\ where\ p_{c_1} = \frac{TP + FN}{TP + FP + FN + TN} * \frac{TP + FP}{TP + FP + FN + TN}$ and $p_{c_2} = \frac{FP + TN}{TP + FP + FN + TN} * \frac{FN + TN}{TP + FP + FN + TN}$), which can also be interpreted as a model's usefulness.
- Sensitivity: Ability to detect true stroke cases. $\frac{TP}{TP + FN}$
- Specificity: Ability to rule out the unimportant class. $\frac{TN}{FP + TN}$

3. Results

3.1 Preparation of data

To prepare the data for analysis, we carefully set each data type and coded variables appropriately. First, the BMI variable had 201 missing values, therefore these observations were removed from the dataset. Second, the gender variable originally had three categories of male, female, and other. However, the other category was only represented by a single observation. Due to the other category not holding statistical significance in the dataset, this observation and category of gender was removed. Therefore, the data analyzed only represents two categories of gender, male and female. Finally, for factor variables with more than two categories, we releveld the factors as to set the appropriate baseline category for the regression models. For the work_type variable, the category of “Never_worked” is considered the baseline and for the Smoking_status variable, the category of “never smoked” is set as the baseline. Therefore, the presence of the alternate categories indicates its contribution to the increase or decrease of the log odds if that category is chosen over the baseline.

In addition to data preparation, there was special consideration in the treatment of the smoking_status variable. In the original dataset, the variable included four categories: formerly smoked, never smoked, smokes and unknown. The unknown data are equivalent to missing values, however removing these values would result in 1483 of the now 4907 observations taken out of the dataset. To retain these values, we employed a Neural Network to impute these unknown smoking statuses by considering remaining both categorical and numerical data. For that data with a smoking status was considered as training data and fit the model. Finally by considering the part of the dataset without smoking status as the test dataset, the model reassigns the smoking status of those individuals to the remaining three categories.

3.1.1 Oversampling

We used a technique of over sampling called SMOTE where we randomly generate new observations that are stroke patients based on the KNN algorithm. We did this until the two classes were even. We can see that we now have 286 non stroke and 286 stroke patients to train our models. This technique should allow our models to see the distinction between the classes again, but we are generating “fake” information so we may be misleading and cause overfitting.

3.1.2 Undersampling Technique

We used a simple technique of rebalancing our class distribution by undersampling the majority class. In our training data we have 143 stroke patients, so by equal weighting we randomly sampled 143 non stroke patients to then have a smaller training dataset, where there are 286 total observations/patients. We then used this smaller dataset to train all of our models. The motivation behind using this technique is if we rebalance the distribution it allows our models to better “see” a distinction between the two classes. However, when it is done some potential information is lost, and we could have had a bad or good sample.

3.2 Analysis

3.2.1 Oversampling

3.2.1.1 Full Logistic Regression model

The first model fitted was the cross-validated full logistic model. The performance metric accuracy, sensitivity, specificity, and Kappa are 0.7731, 0.80303, 0.7769, 0.1783, respectively. This model has about a 77% chance of correctly predicting if a person had a stroke or not. These are fairly good measures, however, this model is difficult to interpret as there are 10 predictors.

$$\begin{gathered} log(\frac{\pi(y)}{1-\pi(y)}) = -5.6536 + 0.2335 gender + 0.08227 age + 1.1916 hypertension + 0.9419 HeartDisease \\ - 0.7620 EverMarried - 13.2281 WorkChildren - 0.1613 WorkGovtJob \\ + 0.1778 WorkPrivate - 0.1953 ResidenceType + 0.0051 AvgGlucoseLevel \\ + 0.0081 bmi + 0.5620 FormerlySmoked + 0.3512 smokes \end{gathered}$$

It is important to note that the WorkSelfEmployment variable is linearly dependent on all other variables in the model, therefore returning an NA coefficient. Therefore, this predictor has been removed for the full logistic regression model on the oversampled training data. At an alpha level of 0.05, the statistically significant variables in the model are age, hypertension, HeartDisease, EverMarried, AvgGlucoseLevel and FormerlySmoked.

3.2.1.2 Stepwise-type variable selection using backwards elimination for logistic regression

The next model is the stepwise variable selection using backwards elimination for logistic regression. The performance metric of accuracy, sensitivity, specificity, and Kappa are 0.5781, 0.95455, 0.56046, 0.0947, respectively. Compared to the full model, the accuracy and kappa are much lower while the trade off between sensitivity and specificity are more unbalanced. The sensitivity is much higher at 0.95 than the specificity which is 0.56.

$$log(\frac{\pi(y)}{1-\pi(y)}) = -5.3396 + 0.0811 age + 1.1901 hypertension + 0.9531 HeartDisease \\ - 0.7063 EverMarried + 0.0060 AvgGlucoseLevel + 0.4945 FormerlySmoked$$

The variables that are statistically significant at an alpha level of 0.05 are age, hypertension, HeartDisease, EverMarried and AvgGlucoseLevel.

3.2.1.3 Ridge logistic regression

The next model is ridge logistic regression. As a reminder, this is another form of variable selection, that minimizes SSE, but adds a penalty term. We have found that our best tuned parameters are alpha equal to 0 and lambda equal to 0.033. The performance metric of accuracy, sensitivity, specificity, and Kappa are 0.7643, 0.75758, 0.76458, 0.1595, respectively.

These results are on par with the performance of the full logistic regression, but has a better balance between sensitivity and specificity.

$$log(\frac{\pi(y)}{1-\pi(y)}) = -4.0373 + 0.2257gender + 0.0515age + 1.0216hypertension + 0.9037HeartDisease \\ - 0.2112EverMarried - 0.7223WorkChildren - 0.0521WorkGovtJob \\ + 0.1166WorkPrivate + 0.1858WorkSelfEmployed - 0.1325ResidenceType \\ + 0.0056AvgGlucoseLevel - 0.0012bmi + 0.5154FormerlySmoked + 0.1904smokes$$

3.2.1.4 Lasso logistic regression

The next model is lasso logistic regression. We have found that our best tuned parameters are alpha equal to 1 and lambda equal to 0.004. The performance metric of accuracy, sensitivity, specificity, and Kappa are 0.7724, 0.78788, 0.77169, 0.174, respectively. It correctly predicts patients that have a stroke at a decent high rate, comparable to that of the full model, with a good kappa value. The balance between sensitivity and specificity are also more balanced for the lasso method than the full logistic regression. The model is reduced to 13 predictors, which makes it slightly more interpretable than the full model. The formula is:

$$log(\frac{\pi(y)}{1-\pi(y)}) = -5.4388 + 0.1947gender + 0.0790age + 1.1260hypertension + 0.8632HeartDisease \\ - 0.5407EverMarried - 0.0675WorkChildren - 0.0873WorkGovtJob \\ + 0.1502WorkPrivate - 0.1325ResidenceType + 0.0049AvgGlucoseLevel \\ - 0.0037bmi + 0.4802FormerlySmoked + 0.2508smokes$$

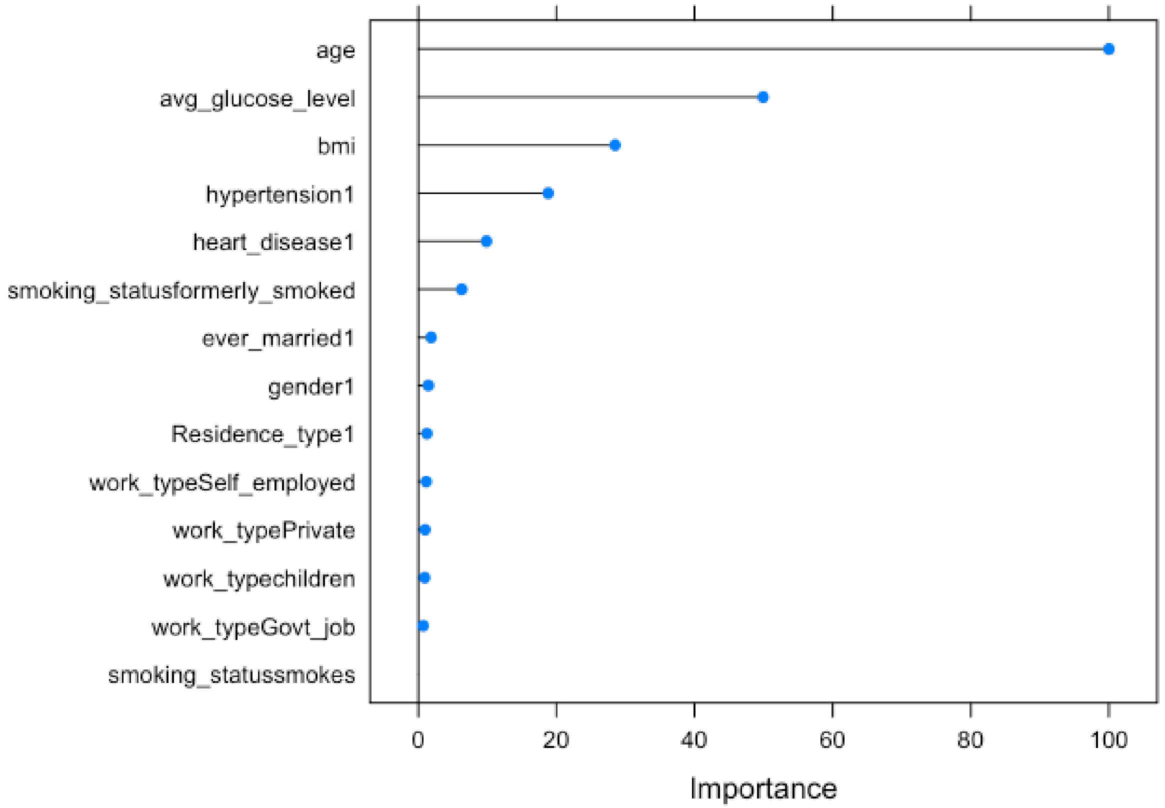
3.2.1.5 Elastic net logistic regression

The next model is elastic net logistic regression. This combines both lasso and ridge regression methods to create a penalty. We have found that our best tuned parameters are alpha equal to 0.25 and lambda equal to 0.059. The performance metrics of accuracy, sensitivity, specificity, and Kappa are 0.7649, 0.77273, 0.76458, 0.1673, respectively. It correctly predicts patients that have a stroke at a decently high rate and has a good balance between the sensitivity and specificity. The performance metrics are slightly less than that of the full model but the method has reduced the model to just 6 predictor variables. Because of this variable reduction, the model is more interpretable than both the lasso and full logistic regression model. The formula is:

$$log(\frac{\pi(y)}{1-\pi(y)}) = -3.5177 + 0.1211gender + 0.04490age + 0.8614hypertension \\ + 0.6997HeartDisease - 0.1419WorkChildren + 0.3320FormerlySmoked$$

3.2.1.6 Random Forest

The final model we trained using SMOTE training data is Random Forest. This is drastically different from logistic regression as this is a tree and splits based on your answer of yes or no. We have 16 predictors thus the optimal subset of variables will be around $\sqrt{16} = 4$. Our performance metrics are accuracy, sensitivity, specificity, and Kappa are 0.6508, 0.89394 0.63940, 0.1157, respectively. A plot of the variables by importance shows which predictors are the most important in classifying the patients to have a stroke or not. The five most important variables are age, average glucose level, bmi, hypertension, and heart disease.



3.2.2 Undersampling

3.2.2.1 Full Logistic Regression model

The first model fitted was the cross-validated full logistic model. As the model is fit and gives us a relatively stable estimate, we then use our test data to see how well it performs. The performance metric accuracy, sensitivity, specificity, and Kappa are 0.716, 0.84848, 0.70982, 0.1441, respectively. The accuracy, sensitivity, specificity, and Kappa are all worse than that of the oversampled training data. There is now only a 71% chance of correctly predicting if a person had a stroke or not compared to the 77% of the full logistic model developed on the oversampled training data. Obviously, we need much better performance if we want to put it into production. Again, this model is difficult to interpret as there are again 10 predictors.

$$
\begin{gathered}
log\frac{\pi(y)}{1-\pi(y)} = -19.37 + -0.1061Gender + 0.08758Age + 0.3469Hypertension + 0.1405HeartDisease \\
- 0.008864EverMarried + 0.4902WorkChildren + 12.87WorkGov + 13.94WorkPrivate \\
+ 13.43WorkSelf + 0.4347Urban + 0.005866AvgGlucoseLevel - 0.01171BMI + 0.9578SmokeFormer \\
+ 0.3840SmokeSmoker
\end{gathered}
$$

The significant predictors at 0.05 significance are age and being a former smoker.

3.2.2.2 Stepwise-type variable selection using backwards elimination for logistic regression

The next model we again use cross validation to train. The model is stepwise or backwards variable selection for logistic regression. The performance metrics of accuracy, sensitivity, specificity, and Kappa are 0.7167, 0.81818, 0.71195, 0.1381, respectively. It correctly predicts patients that have a stroke at a very high rate, but again does not correctly predict the negative class or overall well.. The formula is:

$$
log(\frac{\pi(y)}{1-\pi(y)}) = -6.3196 + 0.0905age + 0.7449WorkPrivate + 0.0051AvgGlucoseLevel + 0.8337FormerlySmoked
$$

The significant variables at $\alpha = 0.05$are age, work_typePrivate, and smoking_statusformerly_smoked. While, AvgGlucoseLevel was significant at $\alpha = 0.10$.

3.2.2.3 Ridge logistic regression

The next model is ridge logistic regression. Reminder that this is another form of variable selection, that minimizes SSE, but adds a penalty. We have found that our best tuned parameters are alpha equal to 0 and lambda equal to 0.074. The performance metric of accuracy, sensitivity, specificity, and Kappa are 0.7133, 0.81818, 0.70839, 0.1359, respectively. It correctly predicts patients that have a stroke at a decent high rate, but does not excel at predicting the negative class or overall well. The model has 15 predictors, so it did not reduce too many variables. Thus, very difficult to interpret. The formula is:

$$\begin{gathered} log(\frac{\pi(y)}{1-\pi(y)}) = -3.4057 + 0.01480 gender + 0.0648 age + 0.3422 hypertension + 0.3997 HeartDisease \\ + 0.3795 NeverMarried - 0.5704 WorkChildren - 0.3719 WorkGovtJob \\ + 0.2463 WorkPrivate + 0.1599 WorkSelfEmployed + 0.2103 ResidenceType \\ + 0.0052 AvgGlucoseLevel - 0.01308 bmi + 0.6173 FormerlySmoked + 0.0901 smokes \end{gathered}$$

3.2.2.4 Lasso logistic regression

The next model is lasso logistic regression. We have found that our best tuned parameters are alpha equal to 1 and lambda equal to 0.042. The performance metric of accuracy, sensitivity, specificity, and Kappa are 0.7058, 0.80303, 0.70128, 0.128, respectively. It correctly predicts patients that have a stroke at a decent high rate, but does not excel at predicting the negative class or overall well. The model has 5 predictors, so it did reduce many variables (15 to 5). This model is very interpretable for that reason. The formula is:

$$\begin{gathered} log(\frac{\pi(y)}{1-\pi(y)}) = -3.9130 + 0.0648 age - 0.1418 WorkGovtJob + 0.01998 WorkPrivate \\ + 0.0022 AvgGlucoseLevel + 0.2503 FormerlySmoked \end{gathered}$$

3.2.2.5 Elastic net logistic regression

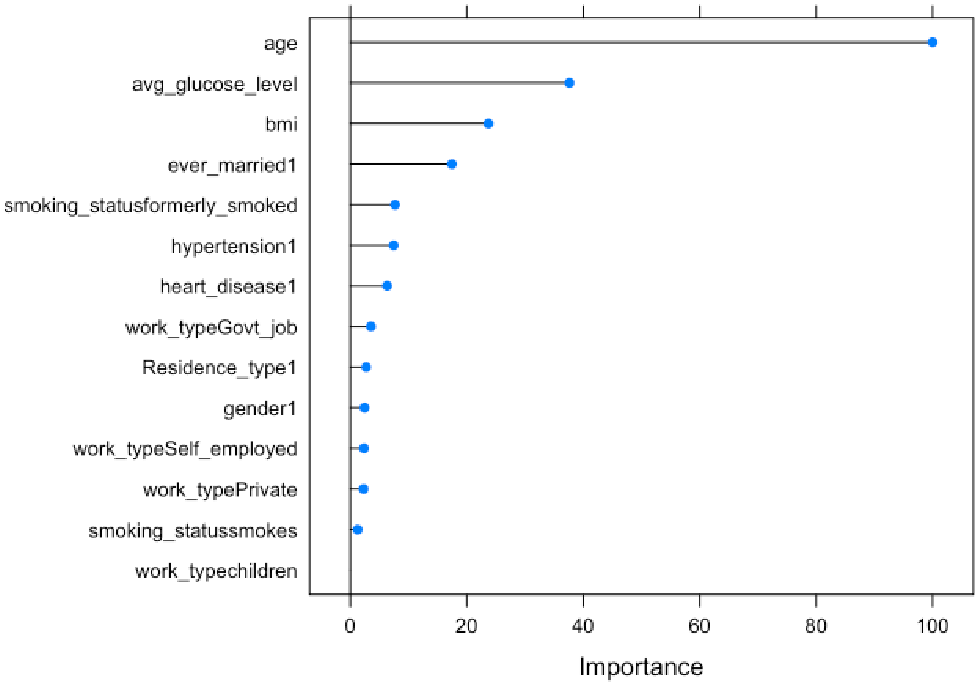
The next model is elastic net logistic regression. We have found that our best tuned parameters are alpha equal to 0.85 and lambda equal to 0.051. The performance metrics of accuracy, sensitivity, specificity, and Kappa are 0.7018, 0.81818 0.6963, 0.1287, respectively. It correctly predicts patients that have a stroke at a decent high rate, but does not excel at predicting the negative class or overall well. The model has 5 predictors, so it did reduce many variables (15 to 5). This model is very interpretable for that reason. The formula is:

$$
\begin{aligned}
log(\frac{\pi(y)}{1-\pi(y)}) &= -3.6381 + 0.0597age - 0.1198WorkGovtJob + 0.0032WorkPrivate \\
&+ 0.0024AvgGlucoseLevel + 0.2458FormerlySmoked
\end{aligned}
$$

3.2.2.6 Random Forest

The final model we trained using our undersample training data is Random Forest. This is drastically different from logistic regression as this is a tree and splits based on your answer of yes or no. We have 16 predictors thus the optimal subset of variables will be around$\sqrt{16} = 4$. Our performance metrics are accuracy, sensitivity, specificity, and Kappa are 0.6549, 0.78788 0.64865, 0.0978, respectively.

A plot of variables by importance is as follows.

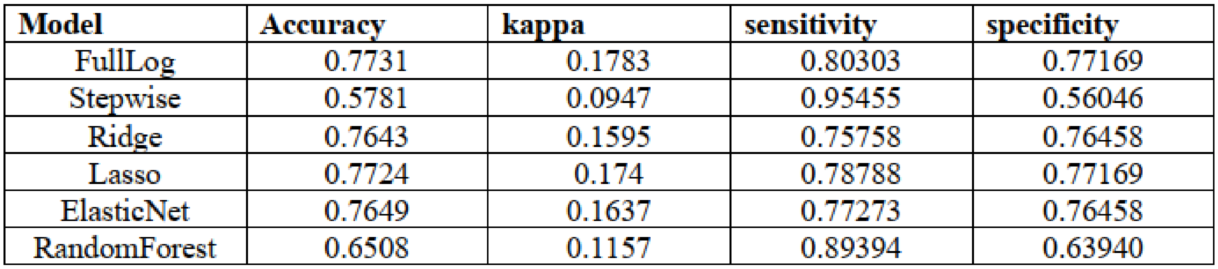


We can see that age is by far the most important predictor age is followed by average glucose level, BMI, and ever married.

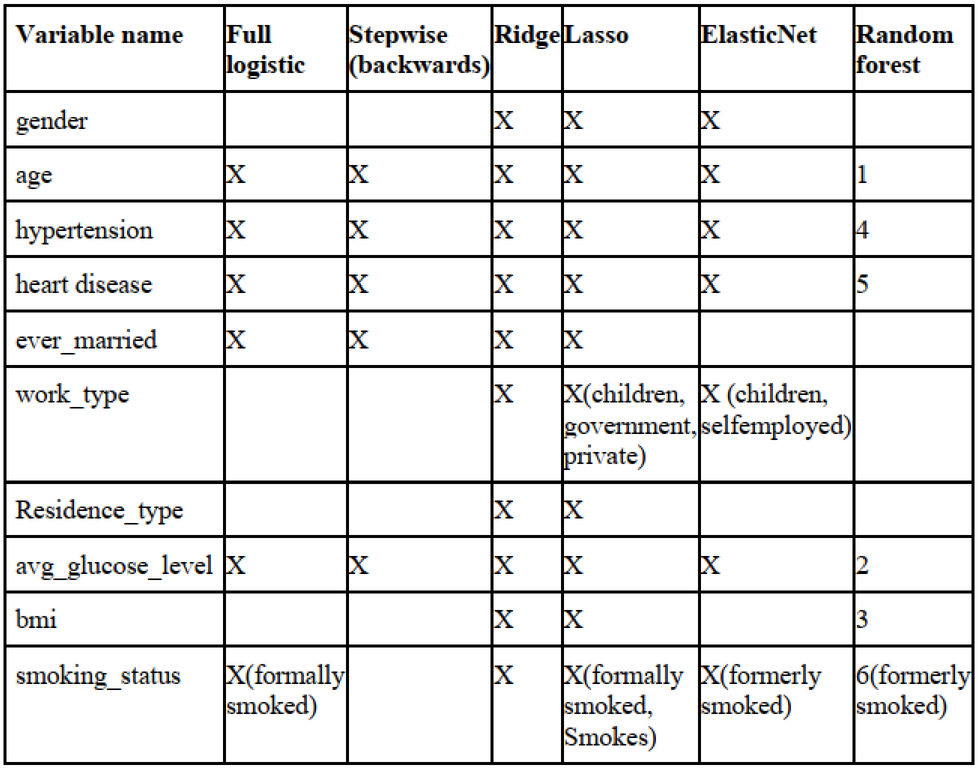
3.3 Discussion

In this section, we are going to discuss all the models we have trained using oversampling and undersampling and compare them within the sampling technique used. At the same time, we will emphasize model performances between these two sampling techniques. Finally, the number of variables used and the most important variables will be considered to focus on prevention of strokes.

Oversampling



Model	Accuracy	kappa	sensitivity	specificity
FullLog	0.7731	0.1783	0.80303	0.77169
Stepwise	0.5781	0.0947	0.95455	0.56046
Ridge	0.7643	0.1595	0.75758	0.76458
Lasso	0.7724	0.174	0.78788	0.77169
ElasticNet	0.7649	0.1637	0.77273	0.76458
RandomForest	0.6508	0.1157	0.89394	0.63940

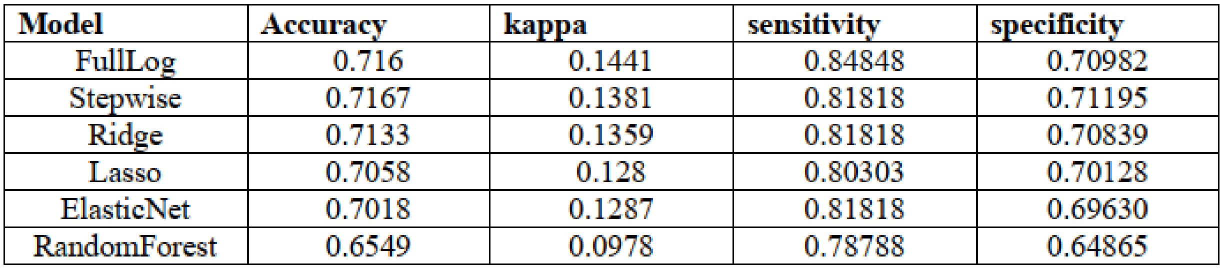


Variable name	Full logistic	Stepwise (backwards)	Ridge	Lasso	ElasticNet	Random forest
gender			X	X	X	
age	X	X	X	X	X	1
hypertension	X	X	X	X	X	4
heart disease	X	X	X	X	X	5
ever_married	X	X	X	X		
work_type			X	X(children, government, private)	X (children, selfemployed)	
Residence_type			X	X		
avg_glucose_level	X	X	X	X	X	2
bmi			X	X		3
smoking_status	X(formally smoked)		X	X(formally smoked, Smokes)	X(formerly smoked)	6(formerly smoked)

In this subsection, we will be discussing models that used the SMOTE training data. The full logistic model had the best performance in the aspects of accuracy, kappa and specificity. Interestingly, oversampled Lasso logistic regression had very close values to the oversampled full logistic model. Also, notice that even the Lasso model is used to do a better job in variable selection as well. In this case, it is fit to all the variables in the model, but drops those with a zero coefficient. This results in some of the categorical variables keeping select categories such as smoking status only being represented by formerly and smokes. Ridge and Elastic Net logistic regression had comparable ratings in accuracy, kappa, and specificity to those models previously mentioned. Even though stepwise backward elimination for logistic regression and random forest had very poor performances in accuracy, kappa and specificity, the highest sensitivity ratings were reported in both of those models.

Apart from model performances, when we try to pay attention to recognize the most important variables, it is very clear that “age” has a huge impact on having a stroke since it appears in every model as an important variable and it is the most important variable in the random forest. Next is the “average glucose level” which also was in all models and the second most important variable from the random forest. And “BMI”, “hypertension” and “heart disease” are also prominent.

Undersampling



Model	Accuracy	kappa	sensitivity	specificity
FullLog	0.716	0.1441	0.84848	0.70982
Stepwise	0.7167	0.1381	0.81818	0.71195
Ridge	0.7133	0.1359	0.81818	0.70839
Lasso	0.7058	0.128	0.80303	0.70128
ElasticNet	0.7018	0.1287	0.81818	0.69630
RandomForest	0.6549	0.0978	0.78788	0.64865

Variable name	**Full logistic**	**Stepwise (backwards)**	**Ridge**	**Lasso**	**ElasticNet**	**Random forest**
gender			X			
age		X	X	X	X	1
hypertension	X		X			6
heart disease			X			7
ever_married			X			4
work_type		X(private)	X	X(government, private)	X(government, private)	
Residence_type			X			
avg_glucose_level			X	X	X	2
bmi			X			3
smoking_status	X(formally smoked)	X (formally smoked)	X	X(formally smoked)	X(formerly smoked)	5(formerly smoked)

In this subsection, we will be discussing models that used the undersampled training data.The full logistic model had the best performance in the aspects of accuracy, kappa, sensitivity and specificity. Interestingly, stepwise backward elimination for logistic regression and ridge logistic regression had close values for all evaluation criterias compared with the full logistic model. Also, the performances of Lasso logistic regression and Elastic net logistic regression are fair ratings in all four metrics. Random forest had the poorest performance, but it is not significantly worse compared to other models.

Apart from model performances, when we pay attention to the importance of variables, it is clear that age has a huge impact on having a stroke, since it appears in every model as an important variable and it is the most important variable in the random forest. Next, average glucose level was included in every model, and the second most important variable from the random forest. BMI, ever married, hypertension, and heart disease are also prominent.

4. Conclusion

Undersampling models generally performed worse than the oversampling models for handling rare events. In the case of our stroke data, the number of observations in the rare class is relatively small at just 143. Therefore, undersampling for our data restricts the dataset to have a small sample size which decreases the performance of the model. However, for oversampling via the SMOTE method, the number of observations in the rare class increases to balance the class representation which improves the accuracy of the classifications.

The top three performing models overall based on simplicity and performance are the stepwise logistic regression model using the undersampled data, the Lasso logistic regression model with oversampling data and the elastic net logistic regression model with oversampling data. Out of the three models, undersampled stepwise logistic regression had the smallest amount of predictors, and the best sensitivity at 0.81818 versus 0.78788 for the other two. However, it was not as accurate as the other two (0.7167 versus 0.7724 and 0.7649). We eliminated that as our “best” model. Since, oversampling lasso and oversampling elastic net had similar metrics we decided that elastic net is our best model. The main reason being that it was simpler with six predictors compared to twelve predictors. Thus, allowing us to better pinpoint factors that contribute to having a stroke. The variables that seemed significant throughout our entire analysis were age, average glucose level, BMI, hypertension and heart disease in the risk of having a stroke.

5. Limitation of the analysis and future direction

5.1 Limitation of the analysis

- ❖ Had lots of missing data, So has to remove all missing data in the BMI variable. Hence lost some information.
- ❖ Missing data imputation for “unknown” was categorized as “formally smoked”, “smokes” and “never smoked” since removing them will reduce the data into a very small dataset.
- ❖ Had to do oversampling and undersampling since the data for individuals having a stroke is very smaller compared with not having stroke.

5.2 Future direction

- ❖ Fitting models using Bagging, Boosting, Neural network etc.
- ❖ Add interaction terms for the models and refit them.
- ❖ Increase the efficiency of the code.

6. References

fedesoriano. (2021, January 26). *Stroke Prediction Dataset*. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Pierce, S. (2019, August 16). *Strokes among younger people on the rise*. TMC News. https://www.tmc.edu/news/2019/05/strokes-among-younger-patients-on-the-rise/#:%7E:text=Smoking%2C%20drinking%20and%20physical%20inactivity,extremely%20important%2C%E2%80%9D%20Gadhia%20said