Stroke Analysis Report

Christine Do, Alyssa Guillory, Johnny Le, Giselle Ruiz, Christopher Turcios, Jimmy Vuong

2022-12-2

Introduction

With heart disease being the leading cause of death in the United States, we think it's important to examine the factors that may lead to strokes. Examining the factors for strokes helps predict and create an image of strokes and heart disease. Our data is sourced from Kaggle with a confidential source and is designated for educational purposes only.

Dataset Information

Data Variables:

- smoking_status: The smoking status of the observation. [Categorical factor w/3 levels]
- id: The unique identifier for the given observation [numeric]
- gender: The gender of the observation [Categorical Factor w/3 levels]
- age: Age of the observation [numeric]
- hypertension: Whether or not the observation has hypertension [Binary dummy variable]
- heart_disease: Whether or not the observation has heart disease [Binary dummy variable]
- ever_married: Whether or not the observation is married [Binary dummy variable]
- work_type: The type of work the observation participates in [Categorical Factor w/5 levels]
- Residence_type: The location in which the observation is located [Categorical factor w/2 levels]
- avg_glucose_level: The average glucose level [numeric]
- bmi: The body mass index [numeric]
- smoking_status: The smoking status of the observation. [Categorical factor w/3 levels]
- stroke: If the observation has had a stroke [Binary dummy variable]

Cleaning the data To use this data set, we first converted variables to their correct representations and then omitted any incomplete observations.

Notable fixes:

- The categorical variable smoking_status was reformatted to remove the level "Unknown", as it was used to represent unavailable data. Any cells that previously had "Unknown" were updated to reflect their unavailable status.
- The quantitative variable bmi was fixed by converting the data type from character to numeric. This change does not apply to non-numeric cells, so the cells containing "N/A" were changed to reflect their unavailable status in the now numeric column.

Main Question

Which factors demonstrate statistical significance in relation to the response variable stroke?

Logistic Regression (Christine Do, Alyssa Guillory, Jimmy Vuong)

We chose to use a logistic regression model for our data because our response variable is qualitative with two classes. It was desirable to have a model predict the probability of a person having a stroke using binary classification. If linear regression had been used to predict the probability of whether or not a person will have a stroke instead, then the model may have predicted Y values outside of our intended range of 0-1. The logistic regression model allows us to enforce this restriction of range.

Model Formula

Our basic logistic regression formula with all our predictors would be as follows:

$$P(stroke = 1) = \frac{exp^{(b0+b1*gender+b2*age+b3*hypertension+b4*heart\_disease+b5*ever\_married+b6*work\_type+b7*Residence\_type+b8*avg\_glucose\_level+b9*bmi+b10*smoking\_status)}}{1 + exp^{(b0+b1*gender+b2*age+b3*hypertension+b4*heart\_disease+b5*ever\_married+b6*work\_type+b7*Residence\_type+b8*avg\_glucose\_level+b9*bmi+b10*smoking\_status)}}$$

This is the logistic regression formula when we consider all of our variables as predictors for the response variable stroke. However, we automatically excluded id as a predictor, as it is a unique number arbitrarily used to identify patients, therefore we know it does not influence our response variable. Using the glm() function, we created our initial logistic model with stroke being the response and all the other variables as predictors.

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2163  -0.3357  -0.1914  -0.1051   3.1357
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.911e+01  4.775e+02  -0.040 0.968076
## genderFemale                7.625e-02  1.681e-01   0.454 0.650091
## genderOther                -1.234e+01  3.956e+03  -0.003 0.997511
## age                         7.304e-02  6.958e-03  10.497  < 2e-16 ***
## hypertension1               5.694e-01  1.828e-01   3.114 0.001845 **
## heart_disease1              3.906e-01  2.211e-01   1.767 0.077249 .
## ever_marriedYes            -1.795e-01  2.627e-01  -0.683 0.494366
## work_typeGovt_job           1.092e+01  4.775e+02   0.023 0.981747
## work_typeNever_worked      -2.958e-01  1.157e+03   0.000 0.999796
## work_typePrivate            1.117e+01  4.775e+02   0.023 0.981339
## work_typeSelf-employed      1.081e+01  4.775e+02   0.023 0.981936
## Residence_typeUrban        -4.830e-03  1.624e-01  -0.030 0.976276
## avg_glucose_level           4.696e-03  1.377e-03   3.409 0.000651 ***
## bmi                         6.501e-03  1.292e-02   0.503 0.614789
## smoking_statusnever smoked -7.476e-02  1.894e-01  -0.395 0.692999
## smoking_statussmokes        3.151e-01  2.305e-01   1.367 0.171639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1141.8  on 3410  degrees of freedom
```

```
## AIC: 1173.8
##
## Number of Fisher Scoring iterations: 16
```

As the summary for health.glm shows, the predictors: age, hypertension, and avg_glucose_level, show high levels of significance and the predictor heart_disease shows a moderate level of significance. Since our p-values for all four of these predictors are less than $\alpha=0.1$, we believe that these are the most important variables in predicting the probability of having a stoke; To confirm that these variables are the most significant, we used the backwards step() function to see if it would give us the same significant variables and it did.

```
#Model created from backwards step()
summary(health2.glm)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
##     family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1420  -0.3357  -0.1927  -0.1072   3.1976
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.632810   0.439480 -17.368  < 2e-16 ***
## age                0.067773   0.006359  10.659  < 2e-16 ***
## hypertension1      0.568379   0.181386   3.134 0.001727 **
## heart_disease1     0.453704   0.216660   2.094 0.036253 *
## avg_glucose_level  0.004701   0.001334   3.524 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1149.9  on 3421  degrees of freedom
## AIC: 1159.9
##
## Number of Fisher Scoring iterations: 7
```

```
#Model created without heart_disease
summary(health3.glm)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + avg_glucose_level,
##     family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0523  -0.3398  -0.1939  -0.1058   3.2054
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.750124   0.437181 -17.728  < 2e-16 ***
## age                0.070099   0.006262  11.194  < 2e-16 ***
```

```
## hypertension1      0.579247    0.180838    3.203 0.001359 **
## avg_glucose_level  0.004932    0.001326    3.720 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1154.1  on 3422  degrees of freedom
## AIC: 1162.1
##
## Number of Fisher Scoring iterations: 7
```

From the health2.glm summary we can see that the AIC is 1159.9, which is an improvement from health.glm's AIC of 1173.8, but not by much. With that in mind, we decided to remove the lowest significant predictor (heart_disease) to perform one last model refitting on our data for further comparison. This time, our AIC for the third model (health3.glm) came out to be 1162.1, which is worse than health2.glm, but still slightly better than health.glm.

$$P(stroke = 1) =$$

$$\frac{exp(-7.632810 + 0.067773*age + 0.568379*hypertension + 0.453704*heart\_disease + 0.004701*avg\_glucose\_level)}{1 + exp(-7.632810 + 0.067773*age + 0.568379*hypertension + 0.453704*heart\_disease + 0.004701*avg\_glucose\_level}$$

Additional information was obtained by using the Goodness of Fit statistical hypothesis test. Given the three models we tested above, we are able to determine each models' null deviance, residual deviance, R^2 value, AIC value, and BIC value. The lower the residual deviance, AIC, and BIC values, the better the model is able to predict the value of the response variable. As for R^2, a value closest to 1 is best.

```
##              Null Deviance Residual Deviance        R^2      AIC      BIC
## health.glm       1411.002          1141.806 0.1907835 1173.806 1272.032
## health2.glm      1411.002          1149.914 0.1850374 1159.914 1190.609
## health3.glm      1411.002          1154.069 0.1820928 1162.069 1186.625

## Train Error Rate Mean:  0.05275447

## Test Error Rate Mean:  0.05080292

## Test Sensitivity Mean:  0.003225806

## Test Specificity Mean:  1
```

Between the three models, health2.glm and health3.glm provided the closest fit for our data, with health2.glm having a slight advantage with lower residual deviance and AIC values. Each model's R^2 value is extremely low, with values being between 0.1 and 0.2, indicating that these models are not a good fit for our data. Even so, we conclude that the best linear regression model for fitting the data and predicting whether or not a stroke will occur is the model health2.glm (predictors: age, hypertension, heart_disease, avg_glucose_level).

Decision Tree (Johnny Le, Giselle Ruiz, Christopher Turcios)

For our second model, we decided to explore using a tree-based model. More specifically, we decided on a classification decision tree, once again due to our response variable being qualitative with two classes. Using a tree-based model would allow us to more easily visualize the probability of a person having a stroke, all based on our predictor variables. Furthermore, with a classification tree, we would be able to perform multiple actions, such as pruning, bagging, and random forest, to improve our model and its' error rate.
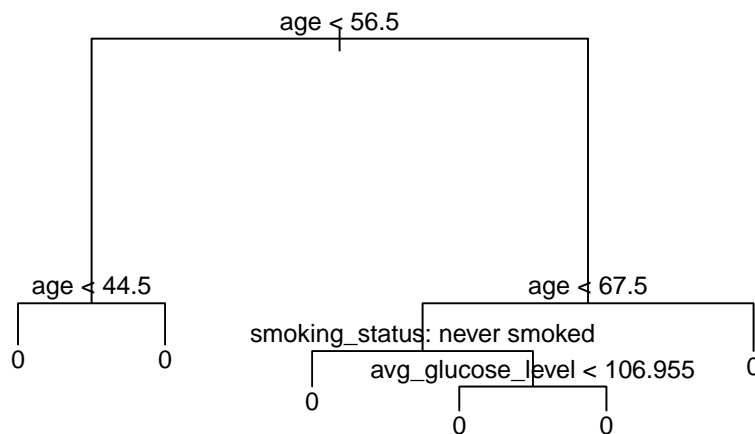
Base Decision Tree

To create a baseline for all of our decision tree based models we created an base decision tree using `tree()` include all data used by our logistic regression model.

Test Error: 0.05107998

## Test Sensitivity: 0.001204819

Test Specificity: 0.9999387

As the summary above shows there is an extremely low Test Sensitivity, leading us to have low confidence in the tree. Rather than the tree having many factors the model chooses a small set of factors which is an uncommon result. To cover all bases we used pruning to determine if it would provide us with a different result.

Pruned Decision Tree

```
## Test Error:  0.05107998
```

```
## Test Sensitivity:  0
```

```
## Test Specificity:  1
```
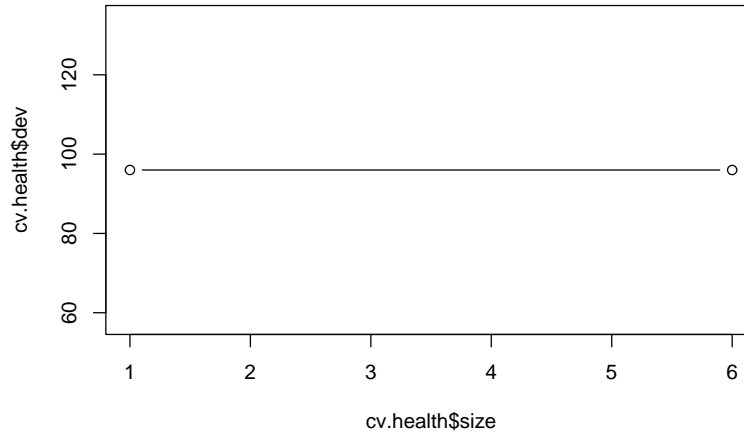

As seen from the summary above the pruned tree provides quite a different result from normal expectations. A Test Sensitivity of 0 would mean that the model entirely fails to predict any occurances of stroke. This would make the pruned model unfit. While this deviates from the norm, it is understable. The base decision tree model had an extremely low Test Sensitivity and any pruning attempts would lower fitment towards our data. As pruning has shown to be unfit for a model we create a bagged tree using `randomForest()`.

Bagged Decision Tree
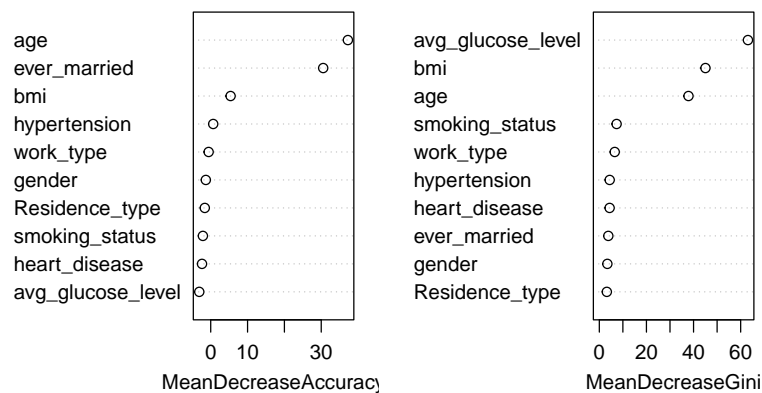
Test Error: 0.05452423

Test Sensitivity: 0.02140718

Test Specificity: 0.9952017
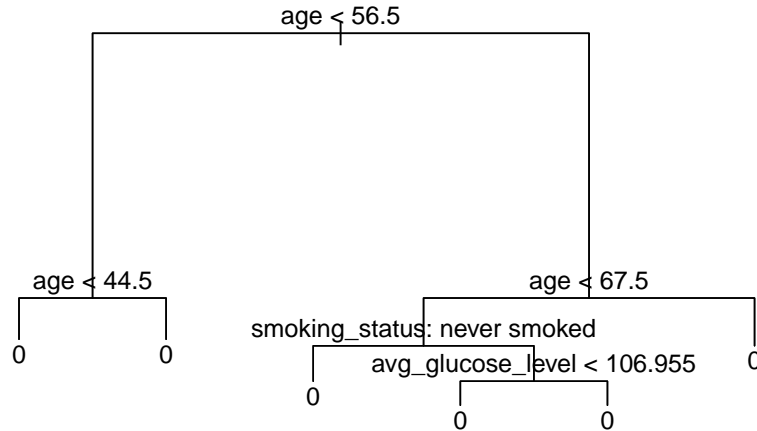
The bagged model provides the best result thus far. However, the model still suffers from an incredible low Test Sensitivity of only 0.02 implying the model can only accurate predict a small portion of strokes.

health.bag

As shown on the importance table BMI has an significant impact on the the occurance of stroke. The decision tree does not reflect this result. Differing from the logisitic model the table indicates low impact from Heart Disease, which contradicts common census that Heart Disease has an impact on stroke.

Random Forest

```
## Test Error:  0.05207239
```

```
## Test Sensitivity:  0.007972732
```

```
## Test Specificity:  0.9985244
```

The Random Forest model provides a slightly different result from our Bagged model. While there is a 0.002 decrease in Test Error this is accompanied by a 0.012 decrease in Test Sensitivity. The decrease in Test Sensitivity is relatively small in scheme of prediction, however, in this situation it makes the Random Forest model worse than our Bagged model.

health.rf
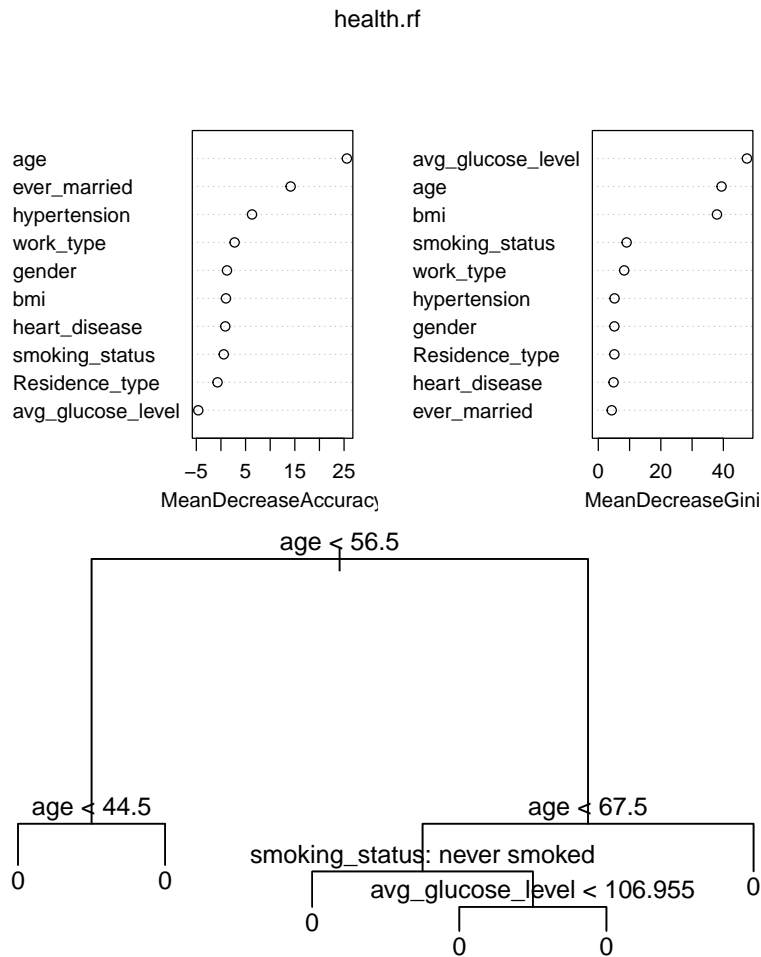
Between all of our models the Bagged model was the best. The model was only the best by a small margin. All the other models performed similarly aside from the pruned model. Each model had a small to zero Test Sensitivity which makes our models impractical for usage.

Our Decision Tree formula would be as follows:

$$stroke \sim age + bmi + avg\_glucose\_level + smoking\_status$$

Conclusion

Overall, the predictors most significant in predicting whether or not a stroke may occur based on our best logistic regression model are: age, hypertension, and avg_glucose_level whereas for our Bagged Tree indicated avg_glucose_level, bmi, age, and smoking_status. Surprisingly, heart_disease was not one of the significant factors for the tree models contrary to what we would have assumed.

Due to the removal of some aspects of the dataset, our models were not able to fit accurately. Removing the rows with NA values resulted in having 250 observations that had a stroke (1) to 180 observations that had a stroke (1). When we fit our models on this cleaned dataset, the 3246 observations that did not have a stroke (0) seems to skew our predictions (there also seems to not be a strong correlation between the predictors for stroke). As a consequence, our models resulted in having values drastically low (MSE) or extremely high (test specificity). The Logistic Regression model outputted an extremely small R^2 value of 0.1850374, indicating that this model is not a good fit for our dataset. Comparatively, the Bootstrap Aggregating model had provided the best overall prediction performance of all decision tree models when looking relatively at the test sensitivity and test specificity. The Bagging model also had a slight advantage over the Logistic Regression model in terms of having a lower MSE value.

Bibliography

Soriano F. (2021, December). Stroke Prediction Dataset Version 1. Retrieved December 2, 2022 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset