Ali Lalani; Kimberly Nguyen; Minh Nguyen; Liam Reilly; Trung Tran

Professor Poliak

MATH 4322

24 April 2025

An Analysis of the Significant Factors that affect Stroke Occurrence with RStudio

I. Introduction

Our project aims to understand the significance of certain factors that affect stroke occurrence for men and women globally. The question which we are asking is this: what variables play the greatest role in stroke occurrence, and at what level do those variables need to be at to bring about a cause of concern for stroke, and do certain variables play a larger or smaller role than thought with today's medical knowledge. The data set we are using is titled "Stroke Prediction Dataset" found on Kaggle. The variables in the data set include id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.

II. Methods

a. Preprocessing

Before constructing our models, we preprocessed the data in the following way. Firstly, we removed all null values in the data set, which were labelled as "Unknown" in the smoking_status variable, and "N/A" for the bmi variable. Of note, after eliminating the rows with null values, the data set shrunk in size from about 5000, to roughly 3000. Afterwards, we encoded the categorical variables, which were gender, ever_married, work_type, Residence_type, and smoking_status, to better optimize the models. This was followed by normalizing the variables with a large standard deviation, which were age and avg_glucose_levels. We then checked for multicollinearity and found nothing, and then checked for outliers. The only outliers present in the dataset were within the avg_glucose_level. After doing some research, we saw that none of the average glucose levels were at an impossible figure, and decided to then not remove those outliers under the impression that removing an aspect of a variable that could play a paramount role in stoke occurrence would cause the model to understate the significance of avg_glucose_level.

b. The Models We Used and Why

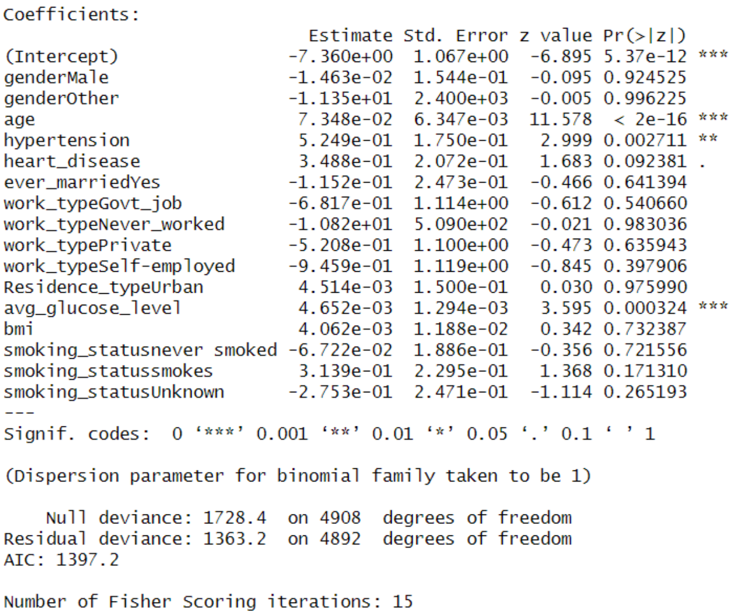
The three models we used for our research were Logistic Regression, Decision Tree, and Random Forest. Logistic Regression is best for binary classification, allowing us to predict stroke occurrence, and it provides probabilities for variables which is useful for understanding how much a factor truly affects stroke occurrence; but, Logistic Regression assumes a linear relationship, meaning it will not accurately reflect a more complex dataset, and it can be sensitive to outliers. We decided to use Decision Tree to capture nonlinear relationships, and for its visually interpretive model; some issues with Decision Tree include overfitting and instability. Lastly, we used Random Forest as an ensemble method to improve the Decision Tree by reducing overfitting, to handle nonlinearity, and to have a visual representation of features importance; some issues though include its black-box nature, the fact that it is expensive computationally, and hyperparameter tuning is needed.

c. Logistic Regression

The Logistic Regression formula is:

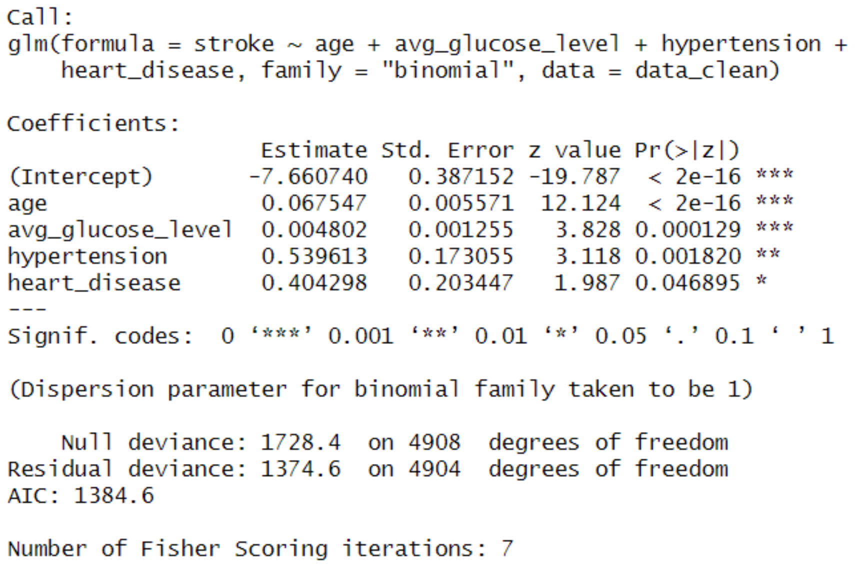
$$p(\text{stroke} = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

We fitted a logistic regression model using all predictors in the dataset. We identified four variables: age, average glucose level, hypertension, and heart disease, as the most statistically significant predictors of stroke due to their lower p-values (p < 0.05). We then refitted a simplified logistic model using these variables to avoid overfitting, lessen complexity, and improve interpretability.



```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -7.360e+00  1.067e+00  -6.895 5.37e-12 ***
genderMale                 -1.463e-02  1.544e-01  -0.095 0.924525    
genderOther                -1.135e+01  2.400e+03  -0.005 0.996225    
age                         7.348e-02  6.347e-03  11.578  < 2e-16 ***
hypertension                5.249e-01  1.750e-01   2.999 0.002711 ** 
heart_disease               3.488e-01  2.072e-01   1.683 0.092381 .  
ever_marriedYes            -1.152e-01  2.473e-01  -0.466 0.641394    
work_typeGovt_job          -6.817e-01  1.114e+00  -0.612 0.540660    
work_typeNever_worked      -1.082e+01  5.090e+02  -0.021 0.983036    
work_typePrivate           -5.208e-01  1.100e+00  -0.473 0.635943    
work_typeSelf-employed     -9.459e-01  1.119e+00  -0.845 0.397906    
Residence_typeUrban         4.514e-03  1.500e-01   0.030 0.975990    
avg_glucose_level           4.652e-03  1.294e-03   3.595 0.000324 ***
bmi                         4.062e-03  1.188e-02   0.342 0.732387    
smoking_statusnever smoked -6.722e-02  1.886e-01  -0.356 0.721556    
smoking_statussmokes        3.139e-01  2.295e-01   1.368 0.171310    
smoking_statusUnknown      -2.753e-01  2.471e-01  -1.114 0.265193    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1363.2  on 4892  degrees of freedom
AIC: 1397.2

Number of Fisher Scoring iterations: 15
```

The final model revealed that age was the greatest predictor: for each additional year, the odds of stroke increased by approximately 7%. Additionally, individuals with hypertension had a 71% increase in stroke odds, and those with heart disease had a 49.8% increase. Higher average glucose levels also slightly increased stroke risk. The model achieved a low Mean Squared Error (MSE = 0.037), indicating that it is a good fit.

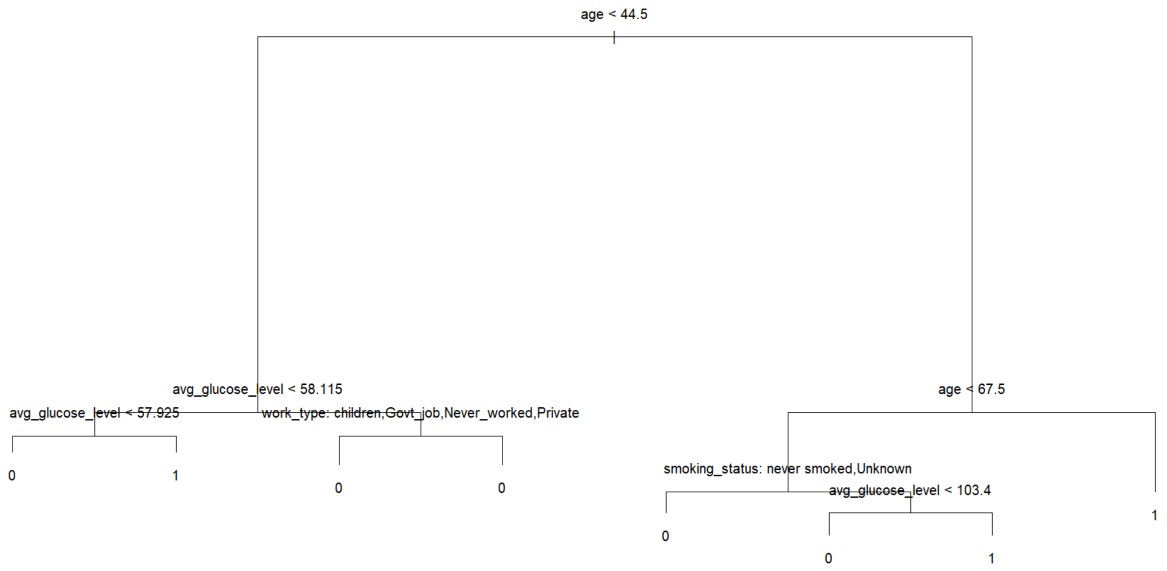


```
Call:
glm(formula = stroke ~ age + avg_glucose_level + hypertension +
    heart_disease, family = "binomial", data = data_clean)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.660740   0.387152 -19.787  < 2e-16 ***
age                0.067547   0.005571  12.124  < 2e-16 ***
avg_glucose_level  0.004802   0.001255   3.828 0.000129 ***
hypertension       0.539613   0.173055   3.118 0.001820 **
heart_disease      0.404298   0.203447   1.987 0.046895 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1374.6  on 4904  degrees of freedom
AIC: 1384.6

Number of Fisher Scoring iterations: 7
```

d. Decision Tree

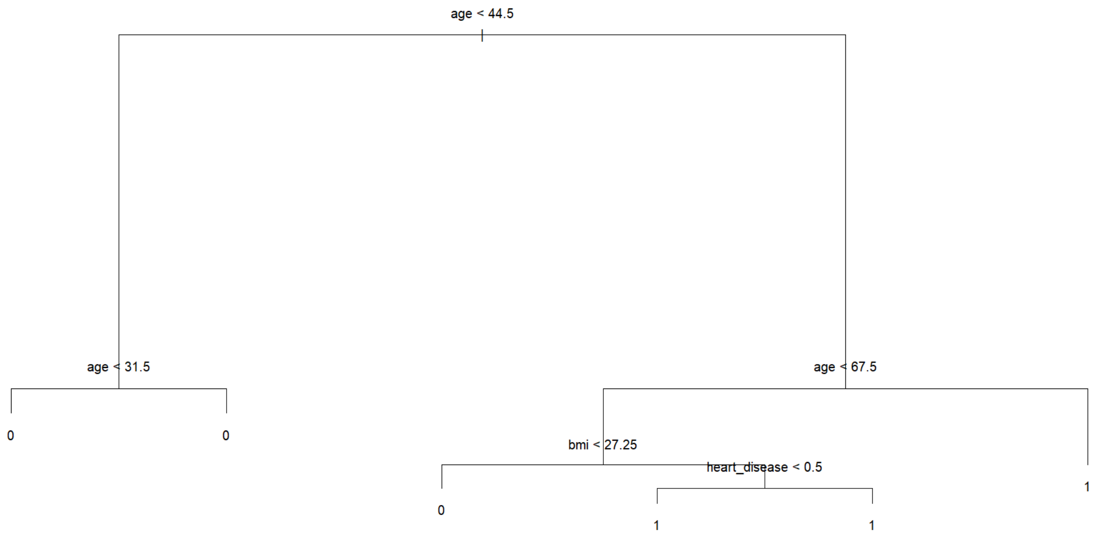
The dataset had significant class imbalance, with only 249 stroke cases compared to 4861 non-stroke cases. This imbalance could lead to potential biased models that favor the majority class and fail to identify stroke cases. To fix this, we applied oversampling to the training data using the ovun.sample() function from the ROSE package. This technique duplicated minority class (stroke = 1) observations until both classes were balanced. Oversampling helped the model better learn patterns associated with stroke cases and improved classification performance for the stroke class. A classification tree was then built to predict stroke using all variables, with oversampling applied to help minimize class imbalance.

The tree prioritized age as the most important variable, making initial splits at 44.5 and 67.5 years old. avg_glucose_level was the second most influential predictor, as it appeared multiple times throughout the tree. Additional predictors included smoking_status and work_type despite their weaker influence. The tree achieved an average misclassification error of 28.8% across 10 randomized train-test splits. While the tree is interpretable and highlights age and glucose levels as key factors in stroke prediction, performance could be improved using pruning or ensemble techniques.


age < 44.5
avg_glucose_level < 58.115
avg_glucose_level < 57.925
work_type: children,Govt_job,Never_worked,Private
0
1
0
0
age < 67.5
smoking_status: never smoked,Unknown
avg_glucose_level < 103.4
0
0
1
1

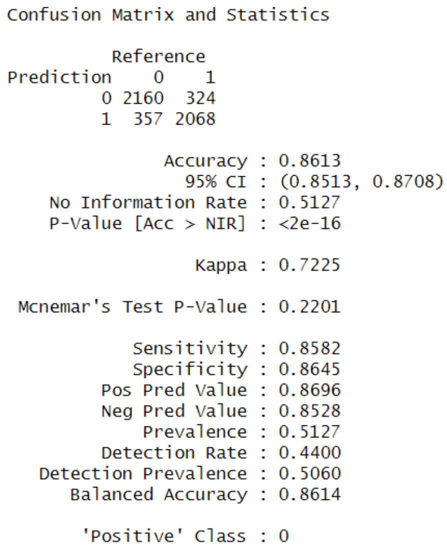
To control model complexity and reduce overfitting, we used cost-complexity pruning. Cross-validation was then applied using cv.tree() to determine the optimal tree size based on misclassification error. The final pruned tree retained only three predictors: age, BMI, and heart disease. This resulted in a simpler, more interpretable model.

The pruned tree achieved a mean test error of 33.99%, slightly higher than the unpruned tree's error of 28.82%, but demonstrated better generalizability. The root split was on age < 44.5, emphasizing age as the strongest predictor. Additional splits showed that higher BMI and presence of heart disease were associated with elevated stroke risk, especially in those over 67.5 years old.



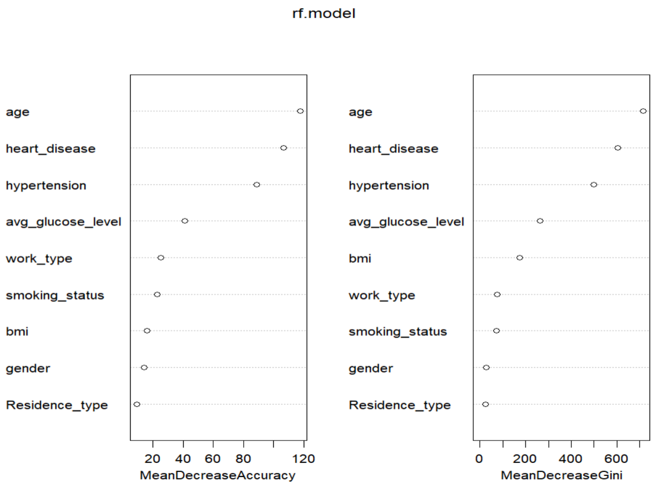
e. Random Forest

We then trained a Random Forest model on a balanced dataset created using the ROSE package. The model achieved an accuracy of 86.13%, with balanced sensitivity and specificity (~86%), and a low Mean Squared Error of 0.09852.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2160  324
         1  357 2068

               Accuracy : 0.8613
                 95% CI : (0.8513, 0.8708)
    No Information Rate : 0.5127
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7225

 Mcnemar's Test P-Value : 0.2201

            Sensitivity : 0.8582
            Specificity : 0.8645
         Pos Pred Value : 0.8696
         Neg Pred Value : 0.8528
             Prevalence : 0.5127
         Detection Rate : 0.4400
   Detection Prevalence : 0.5060
      Balanced Accuracy : 0.8614

       'Positive' Class : 0
```

Variable importance analysis using permutation and Gini index showed us that age, heart disease, and hypertension were the strongest predictors of stroke. This aligns with known medical risk facts and with the results from the logistic regression and decision tree models. While other variables such as bmi, avg_glucose_level, and work_type were included, their relative importance was lower. Overall, the Random Forest model demonstrated strong performance and emphasized consistent predictors of stroke across multiple modeling approaches.



While the Random Forest model had the highest accuracy (86.13%) and performed best in terms of predictive power, the logistic regression model provided the most interpretable results. Depending on the context of the situation, whether for clinical deployment or academic understanding, either model could be considered ideal.

III. Conclusion

All three models had age as the most significant factor, the decision tree splitting at 44.5 years old, signifying that 44.5 appears to be the age that stroke could be a cause of concern. According to the decision tree, individuals below 44.5 were not predicted to have a history of a stroke.

Heart disease appears to be another major factor. According to the logistic regression model, having a history of heart disease increased the odds of stroke by nearly 50%. According to the decision tree, for those individuals between 44.5 to 67.5 and with a BMI greater than 27.25 (or overweight), if those individuals had heart disease, the model predicted a stroke. For the random forest model, it was the second most significant factor.

Hypertension was another significant factor for stroke occurrence. While it didn't play a role in the pruned decision tree, according to the logistic regression model, individuals with a 71% increase in the odds of a stroke. It was also the third leading factor in the decision tree.

The last factor of importance is BMI. While BMI didn't play a major role in the logistic regression model, and it was the 7th most significant factor for the random forest model (7 out of 11), the pruned decision tree did split on BMI after age. Specifically, the decision tree tells us that for individuals from 44.5 to 67.5, if they have a BMI greater than 27.25 (or overweight), the model predicted the individual to have a history of strokes, but for those below 27.25 (or not overweight), the model did not predict the individual to have a history of strokes.

Overall, the most significant factor for strokes appears to be age, according to our decision tree model, age becomes a cause of concern for stroke when individuals are above 44.5. Other major factors at this stage include if the individual has a history of heart disease and hypertension, and if an individual has a BMI of 27.25 (or overweight). Considering that heart disease and hypertrophy can be an outcome from a high BMI, we can conclude that after age, having a BMI of 27.25 or greater is most likely the next most significant factor for predicting stroke occurrence. Afterwards, an individual having heart disease and hypertension along with the being in the previous mentioned categories would greatly increase their odds of a stroke.

Works Cited

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data