KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Stroke and Heart Disease Prediction

Richard Murad, Kennesaw State University
Faculty Advisor: Dr. Marla M. Bell

INTRODUCTION

With stroke and heart disease ranking among the top five leading causes of death in the United States, it is crucial to adopt a data-driven approach to address these public health challenges. Stroke occurs when there is a sudden interruption of blood flow to the brain and heart disease refers to any condition that affects the cardiovascular system. Both tragedies are affecting the core parts of the body, resulting in detrimental consequences. By using advanced statistical modeling and analysis of a 5,109-sample dataset, our goal is to break down the relationships between variables like age, glucose levels, hypertension, smoking status, and gender to predict the likelihood of individuals developing heart disease and stroke. This approach will supply healthcare professionals with the tools needed to make credible decisions for their patients.

METHODS/ASSUMPTIONS

Our methodology involves using SAS code to create two binary logistic models for the response variables (stroke and heart disease). A couple of methods were implemented before running these models. To start with, the 'age' variable was subsetted to exclude patients with unusual decimal values; these values might have been imputed. We are now left with 4,143 patients which is still plenty to work with. Additionally, because there were 11 variables in our dataset, a backwards selection method was utilized to help find better predictors for the response variable in both models; applying a staying level at 0.05 significance enabled us to have interpretable independent variables.

Regarding assumptions, the variance inflation factor was included to assess multicollinearity among all predictor variables for both models. Of course, all the variables needed to be quantitative, so a data step method was necessary to recode the categorical variables (gender and smoking_status) into integers. All the VIFs are under 2 indicating very little multicollinearity among the variables for each respective model.

CODE FOR SAS PROCEDURES

```
*Variable Selection Model 1;
proc logistic data = stroke1;
class smoking_status gender residence_type ever_married;
model stroke = age bmi smoking_status avg_glucose_level heart_disease hypertension gender residence_type
ever_married /selection =b slstay=.05;
run;

*Variable Selection Model 2;
proc logistic data = stroke1;
class smoking_status gender residence_type ever_married;
model heart_disease = age bmi smoking_status avg_glucose_level stroke hypertension gender residence_type
ever_married /selection =b slstay=.05;
run;

*Actual Model 1;
proc logistic data = stroke1 descending plots=(oddsratio(cldisplay=serifarrow) roc);
class hypertension(ref='0');
model stroke = age avg_glucose_level hypertension / lackfit aggregate scale=none ;
run;

*Actual Model 2;
proc logistic data = stroke1 descending plots=(oddsratio(cldisplay=serifarrow) roc);
class smoking_status(ref='never smoked') gender(ref='Female');
model heart_disease = age avg_glucose_level smoking_status gender / lackfit aggregate scale=none;
run;

*Model 1 Assumption Check;
proc reg data = stroke1;
model stroke = age avg_glucose_level hypertension/vif;
run;

*Model 2 Assumption Check;
proc reg data = stroke2;
model stroke = age avg_glucose_level gender_binary smoking_status_dummy/vif;
run;
```

Figure 1: Boxplot distribution of Quantitative Variables

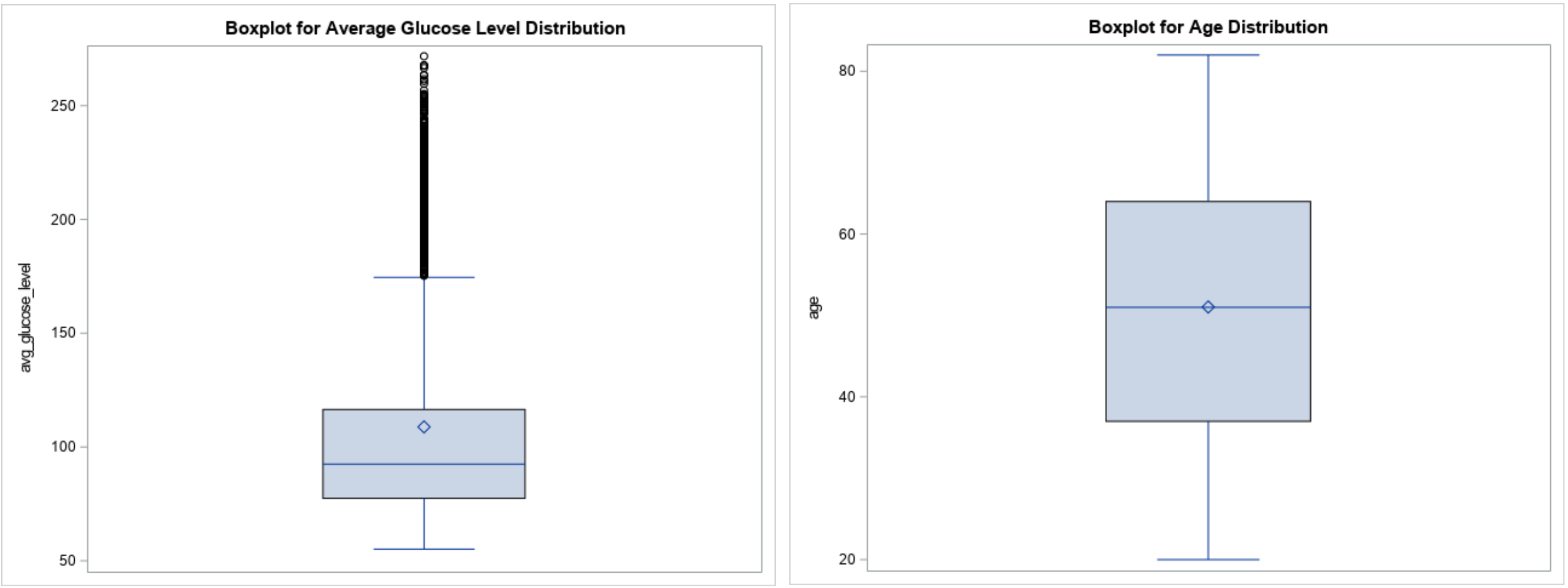

Table 1: VIF Values for Both Models

Stroke Model	
Variable	Variance Inflation
age	1.10399
avg_glucose_level	1.06816
hypertension	1.07732

Heart Disease Model	
Variable	Variance Inflation
age	1.0564
avg_glucose_level	1.05799
gender_binary	1.01107
smoking_status_dummy	1.00456

Figure 2: ROC Curve for Stroke Model (left) and Heart Disease Model (right)

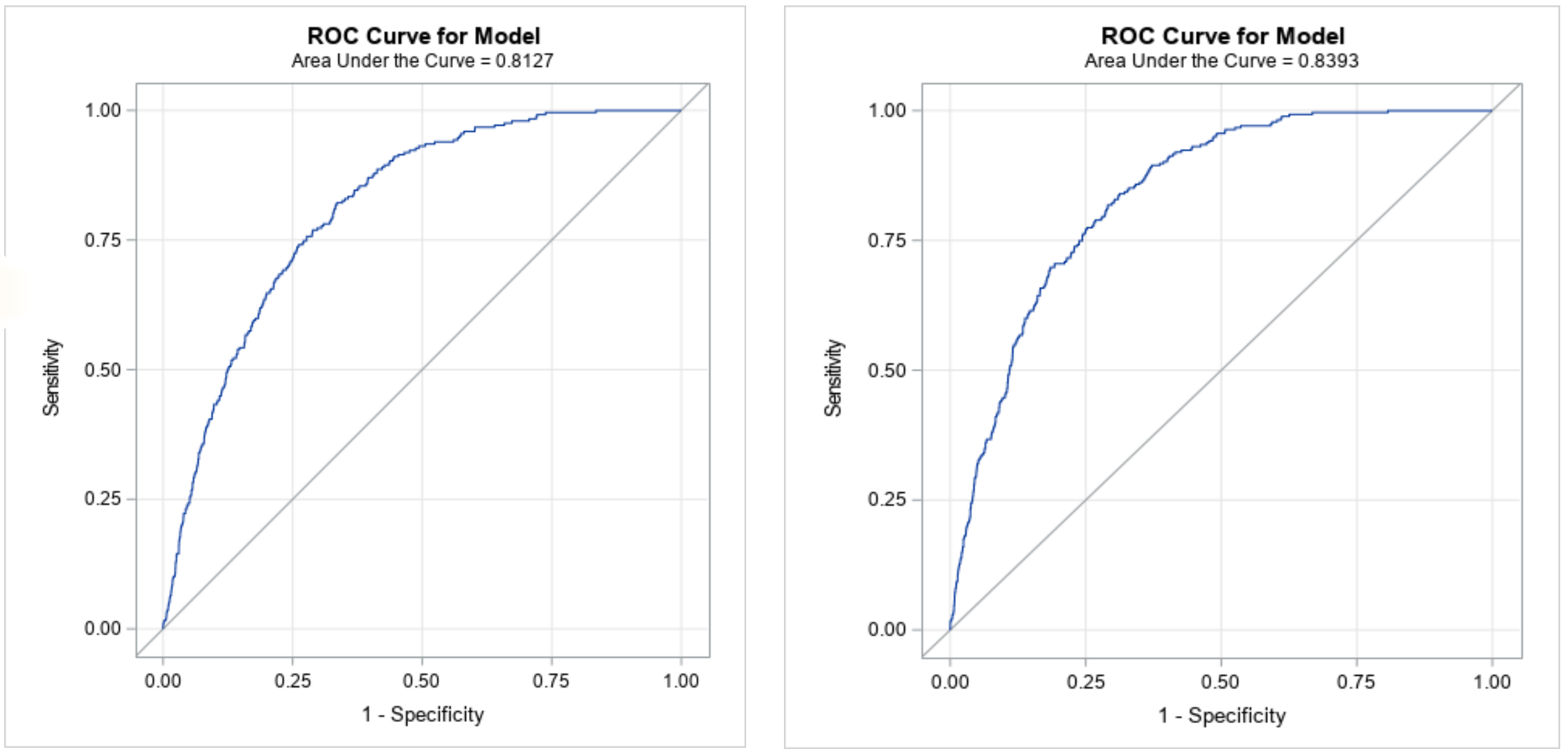

Table 2: Odds Ratio Estimates

Odds Ratio Estimates for Stroke			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.074	1.063	1.085
avg_glucose_level	1.004	1.002	1.007
hypertension 1 vs 0	1.465	1.065	2.014

Odds Ratio Estimates for Heart Disease			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.082	1.071	1.094
avg_glucose_level	1.006	1.004	1.008
smoking_status smokes vs never smoked	2.037	1.41	2.945
gender Male vs Female	2.18	1.668	2.848

RESULTS/INTERPRETATIONS

Stroke Interpretations

An increase in age by one year is associated with an increase in odds of experiencing stroke by approximately 7.4%, given that all other variables are held constant.

An increase in average glucose level in the blood by one milligram per deciliter (mg/dl) is associated with an increase in odds of experiencing stroke by approximately 0.4%, given that all other variables are held constant.

As compared to patients who do not have high blood pressure, patients who do have high blood pressure are associated with an increase in odds of experiencing stroke by approximately 46.5%, given that all other variables are held constant.

Heart Disease Interpretations

An increase in age by one year is associated with an increase in odds of experiencing heart disease by approximately 8.2%, given that all other variables are held constant.

An increase in average glucose level in the blood by one milligram per deciliter (mg/dl) is associated with an increase in odds of experiencing heart disease by approximately 0.6%, given that all other variables are held constant.

As compared to being female, being male is associated with an increase in odds of experiencing heart disease by approximately 118%, given that all other variables are held constant.

As compared to patients who's never smoked, being a smoker is associated with an increase in odds of experiencing heart disease by approximately 103.7%, given that all other variables are held constant.

DISCUSSION

The presented results provide us insights into the odds factors of stroke and heart disease. All these findings make intuitive sense except for the gender variable; a recommended future study is to investigate the reasons why males have significantly higher odds of experiencing heart disease than females. Unlike age and gender, most of these variables are modifiable through lifestyle changes and medical management:

Modifiable Variables

Stroke: Average glucose level and high blood pressure (hypertension)
Heart Disease: Average glucose level and smoking status

By lowering average glucose levels and hypertension, as well as putting an end to smoking, individuals can make positive lifestyle changes and reduce their vulnerability to these serious conditions.