Health Factors' Influence on Stoke Occurrence

Glynn Smith

KU
DEPARTMENT OF
BIOSTATISTICS & DATA SCIENCE
The University of Kansas Medical Center

Table of Contents

Title ..... *- 3 -*
Abstract ..... *- 3 -*
Introduction ..... *- 3 -*
Primary Analysis Aim ..... **- 4 -**
Methods Data Source ..... *- 4 -*
Statistical Analysis ..... **- 4 -**
Preliminary Analysis ..... **- 4 -**
Results ..... *- 5 -*
Limitations ..... *- 6 -*
Discussion ..... *- 6 -*
Conclusion ..... *- 7 -*
References ..... *- 8 -*
Appendix ..... *- 8 -*
Distribution of Variables ..... **- 8 -**
Multicollinearity between Predictor Variables ..... **- 9 -**
Model Selection ..... **- 10 -**
Purposeful Selection ..... - 10 -
Summarizing Predictive Power ..... - 12 -
R-Code ..... **- 13 -**

List of Figures

Figure 1: Box Plots of Numerical Variables ..... - 9 -
Figure 2: Bar Charts for Binary Variables ..... - 9 -
Figure 3: ROC Curve for Prediction of Stroke Occurrence ..... - 13 -

List of Tables

Table 1: Correlation Matrix for the Dataset ..... - 10 -
Table 2: VIF Values for the Final Model ..... - 10 -
Table 3: Purposeful Selection Results ..... - 12 -
Table 4: Classification Table for the Final Model ..... - 13 -

Title

Analyzing the association between health smoking status and stroke occurrence and predicting the occurrence of a stroke based on given health factors.

Abstract

As the 2nd leading cause of death globally, it is crucial to understand the factors that increase the probability of experiencing a stroke.[1] As preventative healthcare becomes more popular, knowledge of these health factors can become the guide for initiatives aimed at reducing one's risk of experiencing a stroke and promoting long-term health and well-being. By identifying and addressing these risk factors, individuals can take proactive steps to safeguard their health and minimize the likelihood of experiencing a stroke, thereby enhancing overall quality of life.

To explore the association between smoking status and the occurrence of a stroke, along with attempting to accurately predict the probability of experiencing a stroke, a final model was built on ambient predictor variables. The final model was built using the purposeful selection approach along with optimizing the Akaike information criterion (AIC) of the model. Surprisingly, smoking status was found to not have a significant impact on the likelihood of experiencing a stroke and was excluded from the final model along with gender. The resulting final model consisted of the age, hypertension, heart disease, and average glucose level variables

Introduction

The data for this analysis utilized through Kaggle.[2] The data source is labeled as 'confidential'. However, upon further investigation, the data set was used in a hackathon for McKinsey Analytics, but I was still unable to find the exact source of the data.[3] There are a total of 5110 patients and the following characteristics were collected:

- ID: unique ID assigned to the patient
- Gender: gender of the patient
- Age: age of the patient
- Hypertension: patient's history of hypertension (high-blood pressure)
- Heart Disease: patient's history of heart disease
- Work Type: patient's job category
- Residence: living environment of the patient
- Average Glucose Level: patient's average glucose level
- BMI: patient's body mass index
- Smoking Status: patient's smoking status
- Stroke: patient's history of experiencing a stroke.

Primary Analysis Aim

The aim of this analysis is to investigate whether or not an individual's history of smoking impacts their risk of experiencing a stroke and to build a model to effectively predict the likelihood of a stroke occurrence based on various health factors. By identifying significant predictors and their respective impact on the risk of a stroke, this analysis aims to give patients insight into potential predictive measures. This could empower individuals to understand how specific health factors contribute to their vulnerability to experiencing a stroke, improving proactive healthcare management strategies.

Methods

Data Source

The dataset consists of 12 variables: id, ranging from 67 to 72940; age, ranging from 0.08 to 82; hypertension, with values 0 or 1; heart_disease, with values 0 or 1; ever_married, with categorical values yes or no; work type, with categorical values of children, govt. job, never worked, private, or self-employed; residence, with categorical values rural or urban; avg_glucose , ranging from 55.12 to 271.74; bmi, ranging from 10.3 to 97.6; smoking_status, with categorical values formerly smoked, never smoked, smokes, and unknown; and stroke, with values 0 or 1.

Statistical Analysis

Data analysis was conducted using the statistical software R and focused on multiple logistic regression. Each predictor variable was individually investigated during the model building process. There were 201 observations identified as having missing values. All of which were BMI variable inputs. Since BMI is a poor indicator of an individual's health, the variable was removed.[6] Furthermore, there was one observation with 'other' in the gender variable and 1544 observations with 'unknown' in the smoking status variable. These observations were removed in an effort to reduce ambiguity in the data. The final model was determined by following the purposeful selection process and a combination of forward and backward selection through the stepAIC() function in R.

Preliminary Analysis

A multiple logistic regression model with all possible predictor variables is considered to further investigate the relationship between the dependent variable and independent variables.

$$logit[P(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

- Y_i represents the occurrence of a stroke for the i^{th} individual, 1 for experiencing a stroke, 0 otherwise.

- X_{i1} is the gender for the i^{th} individual, 1 for male, 0 for female.
- X_{i2} is the age (years) for the i^{th} individual.
- X_{i3} is the hypertension history for the i^{th} individual, 1 for hypertension, 0 otherwise
- X_{i4} is the heart disease history for the i^{th} individual, 1 for heart disease, 0 otherwise.
- X_{i5} is the avg. glucose level for the i^{th} individual.
- X_{i6} is the smoking status i^{th} individual, 1 for never smoked, 0 otherwise.
- X_{i7} is the smoking status i^{th} individual, 1 for currently smokes, 0 otherwise.
- If $X_{i6} = X_{i7} = 0$, then we know the i^{th} individual's smoking status is formerly smoked, as all smoking status coefficients are in reference to an individual whose smoking status is formerly smoked.
- $i = 1,2,3, \dots 3565$
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are the unknown parameters to be estimated.

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the unknown parameters to be estimated.

Results

Model selection was conducted through the purposeful selection method in conjunction with the stepAIC() function in R; resulting in the gender and smoking status variables to be dropped from the preliminary model. A more in-depth analysis is shown in the appendix.

The final model can be expressed as:

$$logit[P(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

- Y_i represents the occurrence of a stroke for the i^{th} individual, 1 for experiencing a stroke, 0 otherwise.
- X_{i1} is the age (years) for the i^{th} individual.
- X_{i2} is the hypertension history for the i^{th} individual, 1 for hypertension, 0 otherwise
- X_{i3} is the heart disease history for the i^{th} individual, 1 for heart disease, 0 otherwise.
- X_{i4} is the avg. glucose level for the i^{th} individual.
- $i = 1,2,3, \dots 3565$

The variable of interest, smoking status, had a very weak positive correlation with the dependent variable and was dropped from the final model along with the gender variable following the purposeful selection process. The final model has a multiple correlation $R = 0.2987$. This value represents the correlation between the observed y values and the

estimated probabilities. The final model has a coefficient of determination $R^2 = 0.1810$. The final model has sensitivity$= 0.7871$, specificity $= 0.7245$, and AUC$= 0.8225$.

Limitations

There were a total of 1545 variables removed from the data set, reducing the size of the original data set by about 30%. The exact source of the data was unable to be located despite the data being used in multiple published analyses and hackathons in the past.[3] The use of the data is reassuring, but not being able to pinpoint the source of the data is troubling. Multiple potential outliers were detected in the variable 'avg. glucose level' were identified from its respective boxplot but were not removed. The age variable appears to be recorded in years and there are inputs with decimal places (0.8, 1.4, etc.) which would indicate a toddler. The work type category 'children' seems to indicate that the patient is a child as only patients 16 and under have this work type category. Furthermore, the work type 'private' indicates that the patient didn't disclose the type of work they are involved in, not meaning the private sector. The variable smoking status has a category 'unknown' which indicates that the smoking status information is unknown for the patient.

Discussion

The estimated regression function from this data analysis is:

$$logit[P(Y = 1)] = -7.5546 + 0.0698X_1 + 0.4765X_2 + 0.3511X_3 + 0.0039X_4$$

- Y_i represents the occurrence of a stroke for the i^{th} individual, 1 for experiencing a stroke, 0 otherwise.
- X_{i1} is the age (years) for the i^{th} individual.
- X_{i2} is the hypertension history for the i^{th} individual, 1 for hypertension, 0 otherwise
- X_{i3} is the heart disease
- X_{i4} is the avg. glucose level for the i^{th} individual.
- $i = 1,2,3, \ldots 3565$

The beta coefficients are the multiplicate effects corresponding to each variable. Each interpretation takes into consideration that every other variable is held constant. The estimated odds of experiencing a stroke is multiplied by $\exp(0.0698) = 1.0723$ for every year increase in a patient's age. The estimated odds of experiencing a stroke is multiplied by $\exp(0.4765) = 1.610428$ if the patient has a history of hypertension compared to that of a patient without a history of hypertension. The estimated odds of experiencing a stroke is multiplied by $\exp(0.3511) = 1.420629$ if the patient has a history of heart disease compared to that of a patient without a history of heart disease.

The estimated odds of experiencing a stroke is multiplied by $\exp(0.0039) = 1.003908$ for every unit increase in average glucose level.

The final model has a relatively poor multiple correlation value $R = 0.2987$ and coefficient of determination $R^2 = 0.1810$. Indicating that the percentage of explained variation of the model is 18.10%, leaving 81.9% of the variation left unexplained. Since these values are a relatively poor metric of predictive power, we turn our focus to the area under the curve (AUC), sensitivity, and specificity, all of which are measures of predictive power. The final model has sensitivity$= 0.7871$, specificity $= 0.7245$, and AUC$= 0.8225$. The model is able to correctly identify 72.45% of the actual negative cases and 78.41% of the actual true cases. The AUC value indicates that the model has good discriminatory power and performs significantly better than randomly guessing. These contradictory values suggest that the model may not explain much of the variance in the response variable in terms of linear correlation, but it is still effective at discriminating between the two classes and correctly classifying instances. This could be due to other factors not captured by the linear model.

This analysis proved surprising in that the gender and smoking status variables were dropped along with the fact that no interaction terms showed signs of being statistically significant. However, I was surprised to find this to be the case as I would have imagined there to be at least one interaction term needed for heart disease, hypertension, and average glucose levels. As it would make sense that hypertension (high blood pressure) and average glucose levels would be correlated as high glucose levels are associated with high blood pressure. Along with the fact that hypertension and high glucose levels are known risk factors for heart disease.[4] Gender being dropped also goes against previous literature that found gender to be correlated with hypertension and high glucose levels.[4] In my opinion, this would warrant further investigation.

Conclusion

Overall, this study exposes the complexity of predicting the likelihood of a stroke. Age, heart disease, hypertension, and average glucose levels were identified as being significant when it came to predicting the likelihood of experiencing a stroke, gender and smoking status went against previous studies and my own hypothesis.[4] These findings, especially regarding the gender variable, warrant further investigation. There were several limitations in the dataset, the main one being the lack of transparency when it came to sourcing the data along with a significant amount of the data being removed for the sake of reducing ambiguity. Further research is suggested with a more thorough approach to data collection and variable consideration. Putting a numerical value on the number of cigarettes someone smokes per day/week would surely provide more insight compared to categorizing an individual's smoking history. Including variables such as exercise frequency and alcohol consumption may also prove insightful.

References

[1] "The top 10 causes of death." *Who.int*, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Stroke prediction dataset. Kaggle; 2023.https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

[3] "McKinsey Analytics Online Hackathon - Healthcare Analytics," *Analyticsvidhya.com*, https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/

[4] CDC, "Women and stroke," *Centers for Disease Control and Prevention*, https://www.cdc.gov/stroke/women.htm.

[5] F. J. He and G. A. MacGregor, "Salt and sugar: their effects on blood pressure," *Pflugers Arch.*, vol. 467, no. 3, pp. 577–586, 2015

[6] CDC, "Body mass index: Considerations for practitioners," *Cdc.gov*, https://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf

[7] Agresti A. (2018). *An Introduction to Categorical Data Analysis*. John Wiley & Sons

Appendix

Distribution of Variables

Figure 1 shows the boxplots of the numerical variables age and average glucose level. Gender, hypertension, heart disease, and smoking status are binary variables; therefore, a bar chart is more appropriate to represent their distribution (figure 2). Figure 2 (b) has a multitude of potential outliers. However, this appears to be real health data and it is entirely possible for a large portion of the sample group to have high average glucose levels. Therefore, I will not be investigating these values further. Figure 1 (a) and Figure 2 (d) appear to be symmetrical. Figures 2 (a), (b), (c), and (e) all have the majority of their distribution in the negative category of the variable.

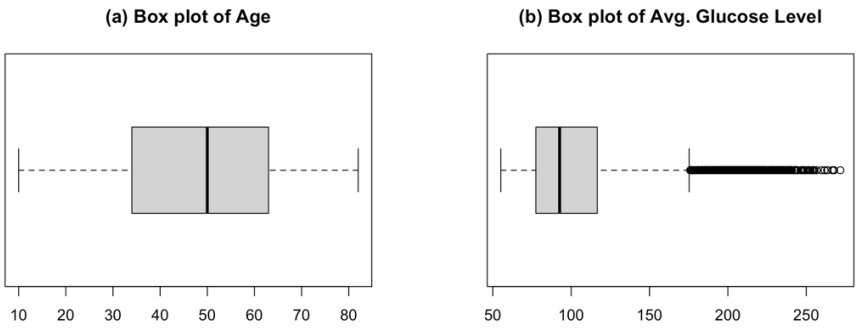

Figure 1: Box Plots of Numerical Variables

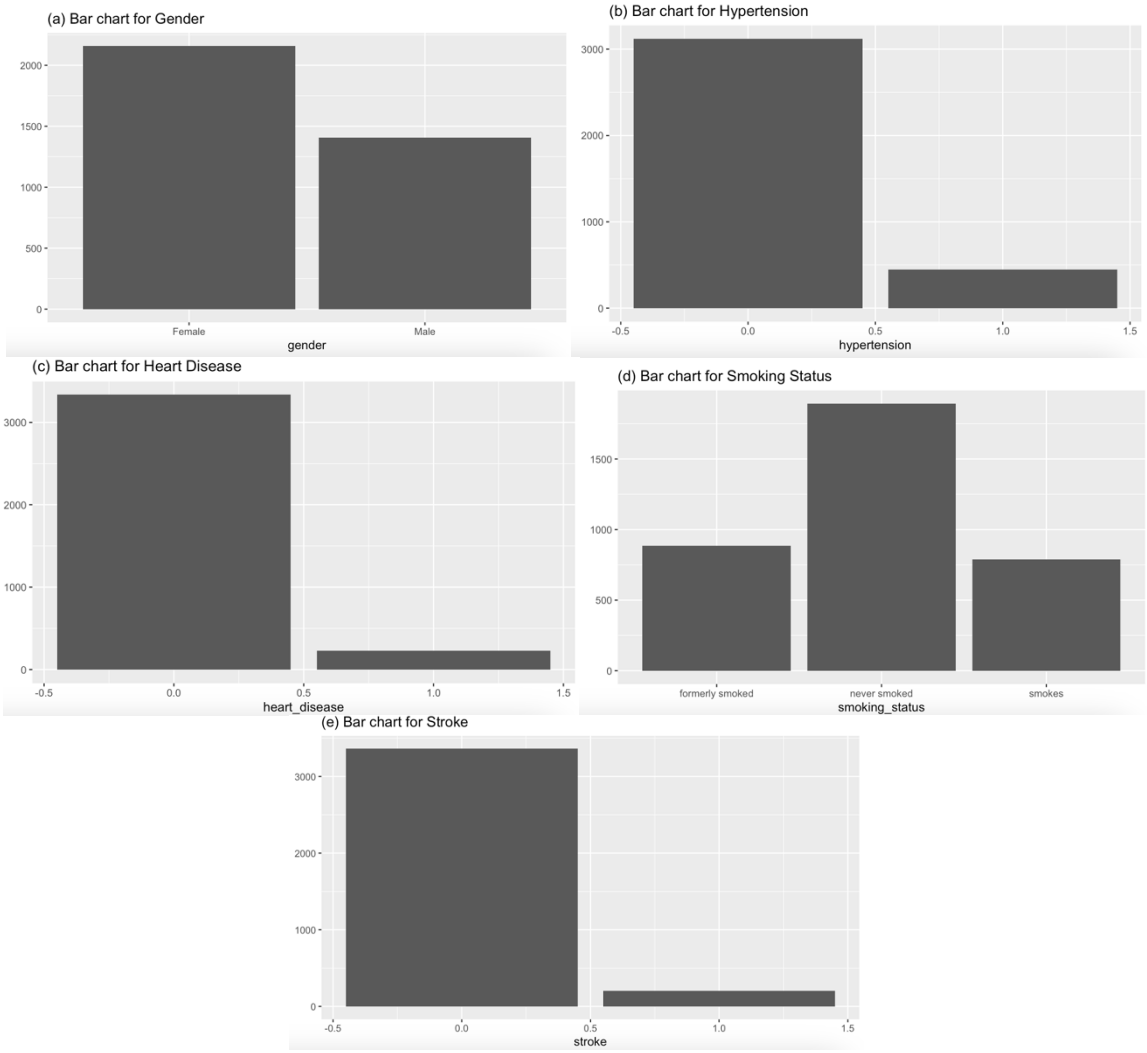

Figure 2: Bar Charts for Binary Variables

Multicollinearity between Predictor Variables

Table 1 shows the Pearson correlation coefficient (r) between all of the variables in the dataset. The closer the values of r are to ± 1, the stronger the correlation is between the

two variables. No variables appear to show any signs of multicollinearity and gender appears to have no correlation whatsoever with any other variable. The variable of interest, smoking status has a r value of 0.0217 between stroke occurrence. A very minuscule positive relationship.

	Gender	Age	Hypertension	Heart Disease	Avg. Glucose Level	Smoking Status	Stroke
Gender	1	0	0	0	0	0	0
Age	0	1	0.2696	0.26427	0.2330	0.0479	0.2508
Hypertension	0	0.2696	1	0.1056	0.1647	-9.3942e-06	0.1347
Heart Disease	0	0.26427	0.1056	1	0.1483	0.0588	0.1293
Avg. Glucose Level	0	0.2330	0.1647	0.1483	1	0.01388	0.1288
Smoking Status	0	0.0479	-9.3942e-06	0.0588	0.01388	1	0.0217
Stroke	0	0.2508	0.1347	0.1293	0.1288	0.0217	1

Table 1: Correlation Matrix for the Dataset

In examining the final model (model 10), the variance inflation factors *(VIF)* were calculated. The *VIF* measures how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables aren't related. When the *VIF* value exceeds 10, it is an indication that multicollinearity is present. Table 2 shows the *VIF* values of the final model, all being less than 10; confirming that multicollinearity is not an issue in the final model.

Age	Hypertension	Heart Disease	Avg. Glucose Level
1.0727	1.0388	1.0538	1.0415

Table 2: VIF Values for the Final Model

Model Selection

Purposeful Selection

As a starting point for selecting explanatory variables, the purposeful selection process was implemented along with AIC selection. Purposeful selection follows four steps:[7]

1. Build an initial main-effects model with key variables and any other variables that show significance as sole predictors (e.g., with P-value < 0.2).
2. Perform backward elimination, retaining variables if they are significant at a stricter level or indicate relevance as a confounder, altering the estimated effect of a key variable when removed.
3. Include variables not in step 1 but that are significant after adjusting for those in step 2, as they might not be directly associated with y but could contribute significantly with other variables present.

4. Test for potential interactions among variables in the model post step 3, employing significance tests at standard levels like 0.05.

Table 3 summarizes the results of fitting and comparing multiple logistic regression models for the occurrence of a stroke, with patient gender *(G)*, age *(A)*, hypertension history *(Ht)*, heart disease history *(Hd)*, average glucose level *(Ag),* and smoking status *(S)* as potential explanatory variables. Each model is symbolized by the included variables. Model 2 *(G)* includes only the gender main effect, whereas model 8 *(A + Ht + Hd +Ag + S),* has age, hypertension, heart disease, and smoking status as main effects. Terms such as "*A*Ht"* represent interaction effects, this one specifically being the interaction effect between age and hypertension.

The likelihood-ratio test is used to compare a more parsimonious model to a more complex model. If a more complex model doesn't fit the data well, it is a strong sign that the selected parsimonious model is appropriate. The likelihood-ratio test has the following hypotheses:

H_0: $The\ more\ parsimonious\ model\ \ is\ sufficient$
H_a: $The\ more\ complex\ model\ sufficient$

If the p-value $> \alpha(0.05)$, then we fail to reject the null hypothesis, otherwise, we reject the null and conclude the alternative.

Six initial main effect models were constructed using the separate individual variables as sole predictors. When tested against the null model, all variables (models: 3, 4, 5, 6, 7), except for the gender variable (model 2) showed signs of being statistically significant as their respective p-values were less than 0.2.[7] Model 8 was then created, including all variables that showed signs of statistical significance, and was then tested against the six initial main effect models. Model 8 was found to be more adequate. Instead of then comparing model 8 to every possible combination of 2, 3, and 4 variables, the stepAIC() function in R was used to find the combination of variables with the lowest AIC. The function starts all variables included in model 8, removing a single variable at each step to maximize the reduction in AIC until a further removal would result in an increase in AIC.[7] AIC judges a model by the distance its fitted values are to the true values. AIC will penalize a model for having more variables and having a smaller AIC is considered to be desirable.

Based on model 8 being the starting point, the stepAIC() function found that model 10 had the lowest AIC value for all possible variable combinations included in model 8. Model 10 has dropped the variable of interest smoking status and the gender variable. Comparing model 10 to model 8, further proved this as the null hypothesis was rejected due to the p-value = $0.2268 > 0.05$.

To see if any cross-product interaction terms are needed, model 11 was created. This model has all 11 possible interaction terms from the variables from model 10 and is the saturated model. Comparing model 11 to model 10 resulted in a p-value =$0.843 > 0.05$, indicating that no interaction terms are needed and that model 10 is adequate.

Model	Explanatory Variables	Deviance	*df*	AIC	Models Compare	Deviance Difference	p-value
1	*None*	1552.1	3564	1554.1			
2	*G*	1551.2	3563	1555.2	(2)-(1)	0.49*(df = 1)*	0.3541
3	*A*	1294.1	3563	1298.1	(3)-(1)	258 *(df=1)*	< 2.2e-16
4	*Ht*	1502.3	3563	1506.3	(4)-(1)	49.8 *(df=1)*	1.69e-12
5	*Hd*	1510.8	3563	1514.8	(5)-(1)	41.3 *(df=1)*	1.301e-10
6	*Ag*	1502.3	3563	1506.3	(6)-(1)	49.8 *(df=1)*	1.766e-12
7	*S*	1541.3	3562	1547.3	(7)-(1)	10.8 *(df=2)*	0.004566
8	*A + Ht + Hd +Ag + S*	1268.2	3558	1282.2	(8)-(3) (8)-(4) (8)-(5) (8)-(6) (8)-(7)	25.9 *(df=5)* 234.1 (*df=5)* 242.6 *(df=5)* 234.1 *(df=5)* 273.1 *(df=3)*	9.353e-05 < 2.2e-16 < 2.2e-16 < 2.2e-16 < 2.2e-16
9	*G + A + Ht + Hd + Ag + S*	1268.2	3557	1284.2	(9)-(8)	0*(df=1)*	0.8602
10	*A + Ht +Hd + Ag*	1271	3560	1281.2	(10)-(8)	3 (*df=2*)	0.2268
11	*A + Ht + Hd + Ag + A*Ht + A*Hd + Ht*Hd + A*Ag + Ht*Ag + Hd*Ag + A*Ht*Hd + A*Ht*Ag + A*Hd*Ag + Ht*Hd*Ag + A*Ht*Ht*Ag*	1264.7	3549	1296.7	(11)-(10)	6.5 (*df=11*)	0.843

Note: G = gender, A = age, Ht =hypertension, Hd = heart disease, Ag = avg. glucose level, S = smoking status

Table 3: Purposeful Selection Results

Summarizing Predictive Power

Table 4 shows a classification table was created to cross classify the binary outcome of y with $\hat{y} = 0$ (absence of a stroke) or $\hat{y} = 1$ (occurrence of a stroke). The prediction for observation i is $\hat{y} = 1$ when $\hat{\pi}_i < \pi_0$, for the sample proportion of patients who experienced a stroke $\pi_0 = \frac{202}{2565} = 0.0567$, otherwise, the prediction for observation i is $\hat{y} = 0$. Sensitivity $= P(\hat{y} = 1|y = 1)$, measures the proportion of actual positive cases that are correctly identified by the model. Specificity $= P(\hat{y} = 0|y = 0)$. is the proportion of actual negative cases that are correctly identified by the model.
In this case, model 10 has a sensitivity $= \frac{159}{202} = 0.7871$ and a specificity $= \frac{2424}{3363} = 0.7208$. The weighted average of sensitivity and specificity $\frac{(159+2424)}{3565} = 0.7245$.

	Prediction, $\pi_0 = 0.0567$	
Actual	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	2424	939
$y = 1$	43	159

Table 4: Classification Table for the Final Model

A receiver operating characteristic (ROC) visualizes the model's sensitivity and specificity at all possible cutoffs of π_0. Analyzing figure 3, with sensitivity on the y-axis and specificity on the x-axis, the ROC plot shows that for a particular value of specificity, better predictive power corresponds to a higher sensitivity. The higher the curve, the better the predictive power.[7] Hence, why we are interested in the area under the curve (AUC). The AUC was found to be 0.8225, a much better value compared to 0.50, which would indicate that predictions were no better than random guessing.

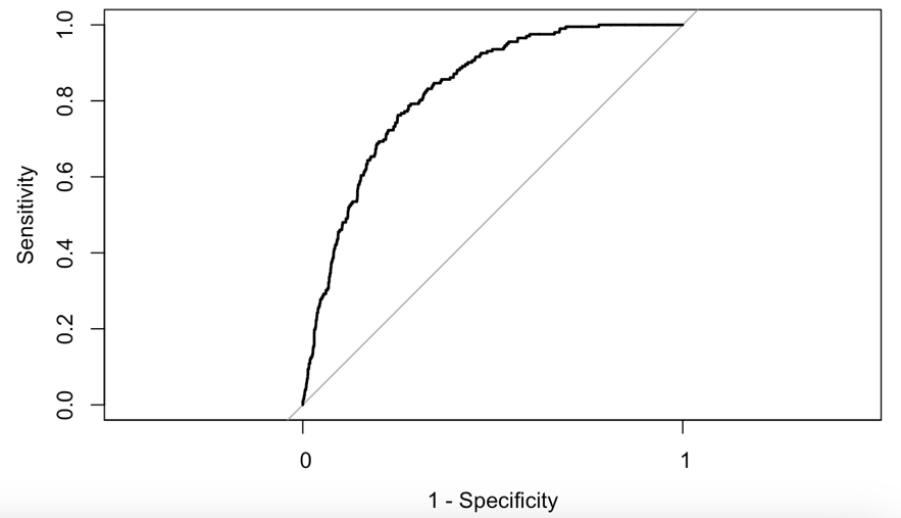

Figure 3: ROC Curve for Prediction of Stroke Occurrence

R-Code

```
#Loading Packages
library(dplyr);library(car);library(leaps);library(MASS);library(bestglm);lib
rary(pROC);library(ggplot2)
###################
#Data exploration
data<-read.csv("~/Documents/Masters/Stat 835/Final Proj/healthcare-dataset-st
roke-data.csv")
nrow(data)#5110
nrow(subset(data, bmi == 'N/A'))#201 N/A values
nrow(subset(data, smoking_status == 'Unknown'))#1544 Unkown smoking status
#Removing unneeded variables
```

```
data<-subset (data ,gender != 'Other')#1 Other variable
data<-subset(data, smoking_status != 'Unknown')
data<-subset(data,select=c(gender, age, hypertension, heart_disease, avg_gluc
ose_level, smoking_status,stroke))
is.na(data)#no missing values
nrow(data)#3565
################
#Purposeful Selection
#Null model
null<-glm(stroke~1, family=binomial, data=data); summary(null)
data %>% ggplot(aes(x=stroke)) + geom_bar() + ggtitle('(e) Bar chart for Stro
ke') + ylab('')

#Gender only model
g.m<-glm(stroke~gender, family=binomial, data=data); summary(g.m)
anova(null, g.m, test = 'LRT')
data %>% ggplot(aes(x=gender)) + geom_bar() + ggtitle('(a) Bar chart for Gend
er') + ylab('')

#Age only model
a.m<-glm(stroke~age, family=binomial, data=data); summary(a.m)
anova(null, a.m, test = 'LRT')#significant/lowest AIC of  1298.1
boxplot(data$age, horizontal =TRUE,main = '(a) Box plot of Age')

#Hypertension only model
h.m<-glm(stroke~hypertension, family=binomial, data=data); summary(h.m)
anova(null, h.m, test = 'LRT')#significant/AIC of 1506.3
data %>% ggplot(aes(x=hypertension)) + geom_bar() + ggtitle('(b) Bar chart fo
r Hypertension') + ylab('')

#Heart disease only model
hd.m<-glm(stroke~heart_disease, family=binomial, data=data); summary(hd.m)
anova(null, hd.m, test = 'LRT')#significant/AIC of 1514.8
data %>% ggplot(aes(x=heart_disease)) + geom_bar() + ggtitle('(c) Bar chart f
or Heart Disease') + ylab('')

#Avg. glucose level only model
ag.m<-glm(stroke~avg_glucose_level, family=binomial, data=data); summary(ag.m
)
anova(null, ag.m, test = 'LRT')#significant/AIC of 1506.3
boxplot(data$avg_glucose_level, horizontal =TRUE,main = '(b) Box plot of Avg.
Glucose Level')

#Smoking Status
ss.m<-glm(stroke~smoking_status, family=binomial, data=data); summary(ss.m)
anova(null, ss.m, test = 'LRT')#significant/AIC of 1547.3
data %>% ggplot(aes(x=smoking_status)) + geom_bar() + ggtitle('(d) Bar chart
for Smoking Status') + ylab('')

#Model w/: age, hypertension, heart disease, avg. glucose, smoking status
m1<-glm(stroke~age+hypertension+heart_disease+avg_glucose_level+smoking_statu
```

```
s,family=binomial,data=data); summary(m1)
anova(a.m, m1,test='LRT');anova(h.m, m1,test='LRT');anova(hd.m, m1,test='LRT'
)
anova(ag.m, m1,test='LRT');anova(ss.m, m1,test='LRT')
#model was more appropriate than single variable models
#full
full.m<-glm(stroke~gender+age+hypertension+heart_disease+avg_glucose_level+sm
oking_status,family=binomial,data=data); summary(full.m)
anova(m1, full.m, test='LRT')#m1 is more appropriate than the full
#Using AIC for model selection
stepAIC(full.m)
test.m <- glm(stroke ~ age + hypertension + heart_disease + avg_glucose_level
,family = binomial, data = data); summary(test.m)
anova(test.m,full.m, test='LRT')#test.m is more appropriate than the full
anova(test.m,m1, test='LRT')#test.m is more appropriate than m1
#testing interaction terms for test.m terms
#Number of 2-way interactions
choose(4,2)
#Number of 3-way interactions
choose(4,3)
#1 4-way interaction
full.sat.m<-glm(formula = stroke ~ age * hypertension * heart_disease * avg_g
lucose_level,family=binomial,data=data)
summary(full.sat.m)
anova(test.m,full.sat.m, test = 'LRT')
#ROC Curve 4.6.2
rocplot<-roc(stroke~fitted(test.m), data=data)
plot.roc(rocplot, legacy.axes = TRUE)

auc(rocplot)
#R and R^2
cor(data$stroke, fitted(test.m)) # 0.2986627
summary(test.m)
1 - (1271.2/1552.1)#adj. R^2: 0.1809806
#multicolinearity
#Temp. modifying data so correlation table can be made. THIS IS ONLY FOR THE
COR TABLE
data$gender <- ifelse(data$gender == "Male", 1, 0)
data$smoking_status <- ifelse(data$smoking_status == "never smoked", 0, ifels
e(data$smoking_status == "formerly smoked", 1, 2))
cor(data); vif(test.m)
#prediction classification table
prop<-sum(data$stroke)/nrow(data); prop #proportion of 1's for stoke: 0.05666
199
predicted <- as.numeric(fitted(test.m)>prop); predicted
xtabs(~data$stroke + predicted)
sensitivity <- 159/(159+43); sensitivity
specificity <- 2424/(2424+939); specificity
prop.cor <-(159+2424)/(2424+939+43+159); prop.cor
```