BST260 Report

Tianxiu Li (Katherine)

Introduction

Background and research question

A stroke occurs when the blood supply to part of the brain is blocked or when a blood vessel in the brain bursts. In either case, a stroke will cause long-term brain damage, disability, or even death [1]. Stroke ranks as the second leading cause of death in the world [2], and more than 795,000 people in the U.S. have a stroke each year [3]. Stroke-related costs, including the cost of health care services, medicines, and missed days of work, are about $53 billion in the U.S. between 2017 and 2018 [3]. High blood pressure, high cholesterol, diabetes, obesity, heart disease, sickle cell disease, age, sex, race and ethnicity, and family history and genetics are major risk factors for stroke [4]. Recognizing the prevalence, complexity, potentially devastating consequences, and social burden of stroke, this study investigates potential risk factors for stroke and whether a patient's likelihood to get a stroke can be predicted by his/her demographic plus clinical features. The results of this study could provide insights for the establishment of earlier and more accurate stroke warnings given a patient's information, which would make the patient be more mindful of his/her health condition so that, ideally, he/she could avoid the worst outcome by seeking immediate help.

Dataset and data preprocessing

This study used the "Stroke Prediction Dataset" downloaded from the Kaggle website [5], which contains 5110 observations among which 249 are participants who had a stroke event and 4861 are healthy controls. Besides the stroke event, this dataset provides information including gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level in the blood, body mass index (BMI), and smoking status for each participant.

Since the goal of the study is to predict the stroke event based on available demographic plus clinical features, the single participant with gender recorded as "other" plus participants with missing data for BMI and smoking status were removed to ensure completeness of the input dataset. After removal, the dataset was left with 3425 observations among which 5.3% had a stroke event. The characteristics of these remaining participants are shown in Table 1. To warrant clinical meaning to this study, three variables – age, average glucose level in blood, and BMI – were redefined and recategorized. Age was broken down into three categories: "<45," "45-65," and ">65." Diabetes status was labeled based on the average glucose level in the blood: "normal" if < 140 mg/dL, "prediabetic" if between 140 mg/dL to 199mg/dL, "diabetic" if ≥ 200 mg/dL [6]. Obesity status was labeled based on BMI: "underweight" if ≤ 18.5 kg/m2, "normal" if between 18.5 kg/m2 and 24.9 kg/m2, "overweight" if between 25 kg/m2 and 29.9 kg/m2, and "obese" if $\geq$ 30 kg/m2 [7].

Exploratory data analysis

After data preprocessing, all predictors for the stroke event become categorical variables for this study. Therefore, an exploratory data analysis was done to see whether certain levels within each predictor are

overly represented in the stroke group (Figure 1). Through visualizing the plots, the stroke group contains more people aged >65, who have hypertension, heart disease, and are. Therefore, age, hypertension, heart disease, and diabetes status could be significant predictors of the stroke event.

Analytical methodology

The outcome variable for this study is the occurrence of the stroke event, which is a dichotomous outcome with levels 0 and 1. Therefore, a logistic regression model is appropriate for modeling this outcome variable. This study fitted two logistic regression models: the full model predicted the stroke event with all the available demographic and clinical information while the simple model predicted the stroke event using only continuous age variable. The simple model was created to comparatively assess the performance of the more complicated full model. After separating the dataset into a training and a test set with a ratio of 8:2, the final model that contains only statistically significant predictors from the full model was fitted to the training set. Since the healthy controls are disproportionally overrepresented in the dataset, a precision versus recall curve (PRC) was generated to determine the threshold probability for assigning the predicted probabilities to the two outcome levels – stroke or healthy control. Following the same steps, the predicted outcomes using the simple model were also assigned with the two outcome levels based on the same threshold probability. Lastly, the performance of the two models was compared in terms of their AIC, accuracy, specificity, and sensitivity.

		Stroke	Control
Total Observation		180	3245
Gender	Female	105	1981
	Male	75	1264
Age Group	<45	5	1426
	45-65	60	1197
	>65	115	622
Obesity Group	Underweight	1	51
	Normal	29	714
	Overweight	64	1029
	Obesity	86	1451
Diabetes Group	Normal	109	2703
	Prediabetes	28	244
	Diabetes	43	298
Hypertension	Yes	57	351
	No	123	2894
Heart Disease	Yes	36	170
	No	144	3075
Ever Married	Yes	160	2439
	No	20	806
Work Type	Children	0	68
	Government Job	23	491
	Private	109	2091
	Self-employed	48	581
	Never Worked	0	14
Residence Type	Rural	86	1594
	Urban	94	1651
Smoking Status	Smokers	39	698
	Formerly Smoked	57	779
	Never Smoked	84	1768

Table 1: Demographic and clinical characteristics of stroke patients and healthy controls

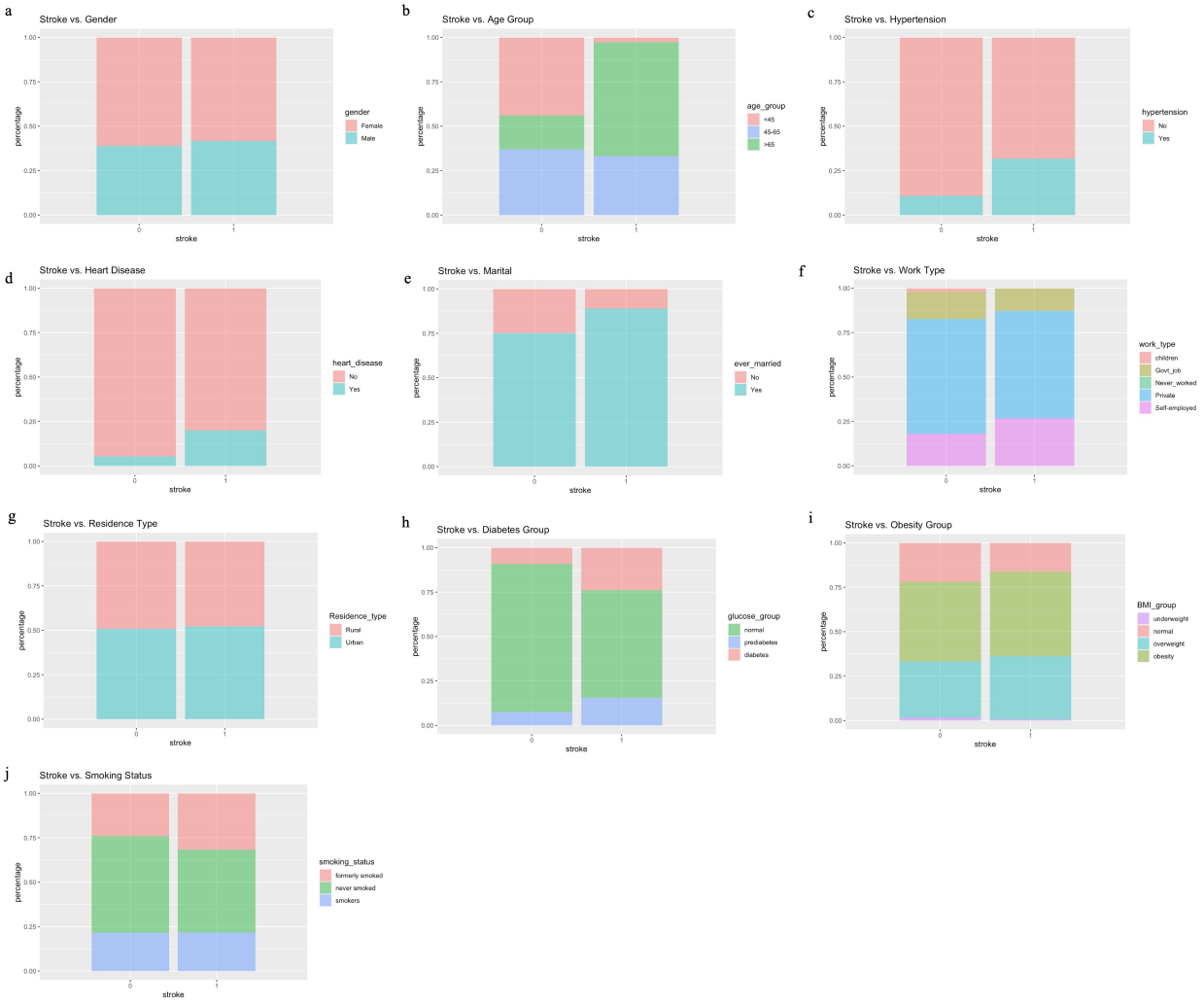

Figure 1: Composition of people within the stroke and healthy control groups categorized by of a) gender, b) age group, c) hypertension, d) heart disease, e) marital status, f) work type, g) residence type, h) diabetes group, i) obesity group, and j) smoking status. The y-axis is the cumulative percentage that adds to 1.

Results

A logistic regression model that predicts the stroke event using age group, gender, hypertension, heart disease, marital status, work type, residence type, diabetes status, obesity status, and smoking status was fitted to the training set. The summary output of this model showed that only age group (45-65: p-value $= 4.520 \times 10^{-6}$; >65: p-value $= 8.850 \times 10^{-12}$), hypertension (p-value $= 2.890 \times 10^{-3}$), heart disease (p-value $= 7.975 \times 10^{-2}$), and diabetes status (prediabetes: p-value $= 8.360 \times 10^{-3}$; diabetes: p-value $= 8.470 \times 10^{-3}$) are statistically significant predictors. Therefore, the final logistic regression model was created by including only these significant predictors. The estimated β coefficients and their corresponding p-values are shown in Table 2.

In addition, this model output suggests that age, hypertension, heart disease, and diabetes are all positively associated with the risk of stroke. More specifically, the odds of stroke for people with age between 45-65 is 11.52 ($e^{2.444}$) times that for people with age <45; the odds of stroke for people with age >65 is 35.59 ($e^{3.572}$) times that for people with age <45; the odds of stroke for people having hypertension is 1.78 ($e^{0.575}$) times that for people not having hypertension; the odds of stroke for people with heart disease is 1.62 ($e^{0.482}$) times that for people without heart disease; the odds of stroke for prediabetic people is 1.92 ($e^{0.652}$) times that for people with normal average glucose level in the blood; and the odds of stroke for diabetic people is 1.73 ($e^{0.549}$) times that for people with normal average glucose level in the blood.

	Estimated β coefficients	P-value
Intercept	-5.760	$< 2.000 \times 10^{-16}$
Age 45-65	2.444	3.320×10^{-6}
Age >65	3.572	7.340×10^{-12}
Hypertension - Yes	0.575	4.390×10^{-3}
Heart Disease - Yes	0.482	4.430×10^{-2}
Prediabetes	0.652	1.266×10^{-2}
Diabetes	0.549	1.395×10^{-2}

Table 2: Final logistic regression model summary

The stroke group is under-represented in the dataset as only 5.3% of the 3425 observations have had a stroke event. Therefore, using a threshold probability of 0.5 to assign predicted probabilities to the two outcome levels – stroke or healthy control – would cause concern and inaccuracy. Consequently, 50 threshold probabilities ranging from 0.01 to 1.00 were tested and their recall and precision are listed in Table 3. Since the goal of this study is to predict the stroke event, instead of pursuing high recall and high precision at the same time, this study emphasized the ability of the selected threshold probability to accurately detect the stroke event. Therefore, a threshold probability of 0.0504 was chosen which detected 30 stroke events out of 38 total stroke events in the test set (Figure 2).

#	p (threshold)	Recall	Precision
1	0.0100	0.428	0.996
2	0.0302	0.430	0.996
3	0.0504	0.708	0.983
4	0.0706	0.802	0.979
5	0.0908	0.802	0.979
6	0.1110	0.906	0.959
7	0.1312	0.906	0.959
8	0.1514	0.906	0.959
9	0.1716	0.960	0.955
10	0.1918	0.969	0.954
11	0.2120	0.969	0.954
12	0.2322	0.969	0.954
13	0.2524	0.980	0.953
14	0.2727	0.989	0.951
15	0.2929	0.995	0.950
16	0.3131	0.995	0.950
17	0.3333	0.995	0.950
18	0.3535	0.995	0.950
19	0.3737	0.998	0.944
20-50	0.3939 - 1.000	1.000	0.945

Table 3: Recall and precision for 50 threshold probabilities

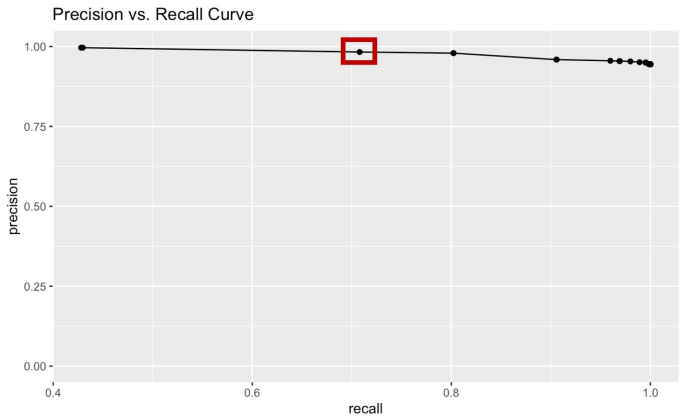

Figure 2: Precision versus recall curve for 50 threshold probabilities. Highlighted point represents the selected threshold probability 0.0504 with recall of 0.708 and precision of 0.983.

To assess the performance of the final model on predicting the outcome of interest, a simpler logistic regression was fitted to the training set. In this model, the stroke event is predicted using only the continuous age variable, which is a statistically significant predictor based on the model summary output (Table 4; p-value $<2.000 \times 10^{-16}$). Moreover, this model output suggests that a 1-unit increase in age is associated with 7.9% ($e^{0.076} - 1 = 0.079$) increase in the odds of stroke, on average. Figure 3 shows that most people who had a stroke event have age greater than 45, and gender and stroke are independent in this study (none of the sex is differentially represented in the stroke or the control group). Moreover, the prediction curve has a sigmoidal shape that matches the trend of probabilities depicted by a logistic regression model.

	Estimated β coefficients	P-value
Intercept	-7.411	$< 2.000 \times 10^{-16}$
Age	0.076	$< 2.000 \times 10^{-16}$

Table 4: Simple logistic regression model summary

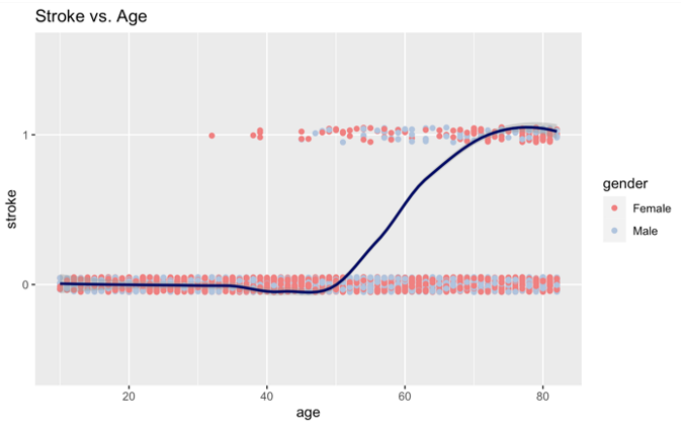

Figure 3: Stroke versus age with prediction curve. Data points are colored based on gender. The shaded gray area displays the confidence interval.

Compared to the simple model, the final logistic regression model has a lower AIC score. When using the same threshold probability of 0.0504, the final model has higher accuracy, sensitivity, and specificity. Breaking down those statistics, the final model successfully detects 30 out of 38 stroke events and 458 out of 647 healthy controls in the test set. However, the simple model only successfully detects 27 out of 38 stroke events and 451 out of 647 healthy controls in the test set (Table 5). Therefore, the full model using age group, hypertension, heart disease, and diabetes status as predictors performs better in predicting the stroke event.

	Final logistic regression model	Simple logistic regression model
AIC	929.86	937.27
Accuracy	0.7124	0.6978
Sensitivity	0.7079	0.6971
Specificity	0.7895	0.7105
# Correctly predicted stroke outcome	30 (38)	27 (38)
# Correctly predicted healthy outcome	458 (647)	451 (647)

Table 5: Performance comparison between the final and simple logistic regression model

Conclusion

This study explores potential risk factors for stroke and investigates whether a patient's likelihood to get a stroke can be predicted by these factors. Through fitting a logistic regression model, age group, hypertension, heart disease, and diabetes status were found to be statistically significant predictors and associated with a higher risk of stroke. To assess the predictive performance of the model via machine learning, a threshold probability of 0.0504 was selected based on the precision versus recall curve plus the goal of the study, and a simple logistic regression model with continuous age being the only predictor was created for comparison purposes. Predicting the stroke events using these two algorithms shows that the more complex model has higher accuracy, specificity, and sensitivity, and thus this model is favored over the simple model.

The analysis was successful as the identified risk factors are consistent with those suggested by the Centers for Diseases Control and Prevention (CDC) [4] and the accuracy of the predictive model is acceptable. However, this study has several limitations. First, the stroke group is under-represented in the dataset. Thus, the available demographic and clinical data were not balanced between the two groups. The high prevalence of the healthy controls would impair the robustness of the model's accuracy. Second, to account for group imbalance, this study manually chose a threshold probability based on the precision versus recall curve. Yet, this would weaken the generalizability of this study as this certain threshold probability may work well only for this specific dataset. Third, this study did not test the association between each pair of categorical predictors. For instance, hypertension is a predominant risk factor for cardiovascular disease [8]. Therefore, a more parsimonious logistic regression model could be viable. Lastly, the dataset does not include information on other critical risk factors for stroke such as race, ethnicity, and family history. Thus, the model still lacks the ability to capture all possible predictors for the stroke event.

A future study should base on a more thorough literature review, from which past prediction methodologies and future directions could be referenced. Moreover, a more group-balanced dataset could be used in future studies. Alternatively, applying a method to control the prevalence would be critical to make the accuracy of the model informative and convincing. Lastly, exploring a way to better graph logistic regression models would help ameliorate data visualization of the study.

References

1. "About Stroke." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 2 Nov. 2022, www.cdc.gov/stroke/about.htm.
2. Murphy, Stephen JX., and David J. Werring. "Stroke: Causes and Clinical Features." Medicine, vol. 48, no. 9, 2020, pp. 561–566., doi:10.1016/j.mpmed.2020.06.002.
3. "Stroke Facts." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 14 Oct. 2022, www.cdc.gov/stroke/facts.htm.
4. "Know Your Risk for Stroke." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 Apr. 2022, www.cdc.gov/stroke/risk_factors.htm.
5. Fedesoriano. "Stroke Prediction Dataset." Kaggle, 26 Jan. 2021, www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.
6. "Diagnosis." American Diabetes Association, diabetes.org/diabetes/a1c/diagnosis.
7. American Diabetes Association Professional Practice Committee. "8. Obesity and Weight Management for the Prevention and Treatment of Type 2 Diabetes: Standards of Medical Care in Diabetes-2022." American Diabetes Association, American Diabetes Association, 16 Dec. 2021, diabetesjournals.org/care/article/45/Supplement_1/S113/138906/8-Obesity-and-Weight-Management-for-the-Prevention.
8. Tackling, Gray, and Mahesh B. Borhade. "Hypertensive Heart Disease." National Library of Medicine, 2022, www.ncbi.nlm.nih.gov/books/NBK539800/.

Appendix: all code for this report

```
######### Library #########
library(tidyverse)
library(ggplot2)
library(caret)
library(dbplyr)
library(psych)

######### Read in data #########
dat<-read.csv("stroke.csv")

######### Data cleaning #########
# left with 3425 observations
dat<-dat|> filter(gender!="Other", bmi!="N/A", smoking_status!="Unknown")
dat|>
  group_by(stroke)|>
  summarise(count=n())

######### Data preprocessing #########
dat<-dat|>
  mutate(age_group=case_when(
  age<45~"<45",
  age>=45&age<=65~"45-65",
  age>65~">65"))|>
  mutate(BMI_group=case_when(
  as.numeric(bmi)<=18.5~"underweight",
  as.numeric(bmi)>=18.6&as.numeric(bmi)<=24.9~"normal",
  as.numeric(bmi)>=25&as.numeric(bmi)<=29.9~"overweight",
  as.numeric(bmi)>=30~"obesity"
  ))|>
  mutate(glucose_group=case_when(
  avg_glucose_level<140~"normal",
  avg_glucose_level>=140&avg_glucose_level<200~"prediabetes",
  avg_glucose_level>=200~"diabetes"
  ))|>
  mutate(age_cat=case_when(
  age<45~0,
  age>=45&age<=65~1,
  age>65~2))|>
  mutate(BMI_cat=case_when(
  as.numeric(bmi)<=18.5~0,
  as.numeric(bmi)>=18.6&as.numeric(bmi)<=24.9~1,
  as.numeric(bmi)>=25&as.numeric(bmi)<=29.9~2,
  as.numeric(bmi)>=30~3
  ))|>
  mutate(glucose_cat=case_when(
  avg_glucose_level<140~0,
  avg_glucose_level>=140&avg_glucose_level<200~1,
  avg_glucose_level>=200~2
  ))|>
  mutate(smoke_cat=case_when(
  smoking_status=="never smoked"~0,
```

```
  smoking_status=="formerly smoked"~1,
  smoking_status=="smokers"~2
  ))

######### EDA #########
a=sum(dat$stroke==0)
b=sum(dat$stroke==1)

# Gender
gen<- dat|>
        group_by(gender,stroke)|>
        summarise(count=n())
gen|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=gender))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Gender")

# Age
age<- dat|>
        group_by(age_group,stroke)|>
        summarise(count=n())
age|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=age_group))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Age Group")+
  scale_fill_discrete(breaks=c("<45","45-65",">65"))

# Hypertension
hyper<- dat|>
        group_by(hypertension,stroke)|>
        summarise(count=n())
hyper|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=hypertension))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Hypertension")

# Heart disease
heart <-dat|>
        group_by(heart_disease,stroke)|>
        summarise(count=n())
heart|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=heart_disease))+
```

```
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Heart Disease")

# Marital
marital<-dat|>
        group_by(ever_married,stroke)|>
        summarise(count=n())
marital|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=ever_married))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Marital")

# Work type
work<-dat|>
        group_by(work_type,stroke)|>
        summarise(count=n())
work|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=work_type))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Work Type")

# Residence type
resi<-dat|>
        group_by(Residence_type,stroke)|>
        summarise(count=n())
resi|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=Residence_type))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Residence Type")

# Average glucose level
glu<- dat|>
        group_by(glucose_group,stroke)|>
        summarise(count=n())
glu|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=glucose_group))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Diabetes Group")+
  scale_fill_discrete(breaks=c("normal","prediabetes","diabetes"))
```

```r
# BMI
BMI<- dat|>
        group_by(BMI_group,stroke)|>
        summarise(count=n())
BMI|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=BMI_group))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Obesity Group")+
  scale_fill_discrete(breaks=c("underweight","normal","overweight","obesity"))

# Smoking status
smok<-dat|>
        group_by(smoking_status,stroke)|>
        summarise(count=n())
smok|>mutate(percentage=case_when(
  stroke==0~count/(a),
  stroke==1~count/(b)))|>
  ggplot(aes(stroke,percentage,fill=smoking_status))+
  geom_bar(stat="identity",alpha=0.5)+
  scale_x_discrete(limits=c(0,1))+
  ggtitle("Stroke vs. Smoking Status")

######### Create training and test set #########
set.seed(1)
test_index <- createDataPartition(dat$stroke,times=1,p=0.2,list=FALSE)
test_set <- dat[test_index,]
train_set <-dat[-test_index,]

######### Prediction #########
# Fit Logistic Regression
summary(glm(stroke~factor(age_cat)+factor(gender)+factor(hypertension)+factor(heart_disease)+factor(eve

# Final Model
summary(glm(stroke~factor(age_cat)+factor(hypertension)+factor(heart_disease)+factor(glucose_cat),data=

# Prediction
glm_fit <- train_set|>glm(stroke~factor(age_cat)+factor(hypertension)+factor(heart_disease)+factor(gluc
p_hat_logit<- predict(glm_fit,newdata=test_set,type="response")

# Determine the threshold
probs<-seq(0.01, 1, length.out = 50)
PRC<-map_df(probs, function(p){
y_hat_logit<- ifelse(p_hat_logit>p,1,0)|>factor()
list(p=p,
     recall=confusionMatrix(y_hat_logit, factor(test_set$stroke))$byClass[["Sensitivity"]],
     precision=confusionMatrix(y_hat_logit, factor(test_set$stroke))$byClass[["Pos Pred Value"]])})
PRC

# Precision and recall curve
PRC|>
```

```
  ggplot(aes(recall,precision))+
  geom_point()+
  geom_line()+
  ylim(c(0,1))+
  ggtitle("Precision vs. Recall Curve")

# choose p=0.0504 as the cutoff point
y_hat_logit<- ifelse(p_hat_logit>0.0504,1,0)|>factor()
confusionMatrix(y_hat_logit, factor(test_set$stroke))

# Fit simple model
mod_simple=glm(stroke~age,data=train_set,family="binomial")
summary(mod_simple)

# Use p=0.0504 as the cutoff point
p_hat_simple<-predict(mod_simple,newdata=test_set,type="response")
y_hat_simple<-ifelse(p_hat_simple>0.0504,1,0)|>factor()
confusionMatrix(y_hat_simple, factor(test_set$stroke))

# Prediction curve
y_hat_simple2<-ifelse(p_hat_simple>0.0504,1,0)
age_plot<-dat|>
  ggplot(aes(age,stroke,col=gender))+
  geom_jitter(width=0,height=0.05)+
  scale_color_manual(values=c("lightcoral","lightsteelblue"))+
  ggtitle("Stroke vs. Age")+
  scale_y_discrete(limits=c(0,1))
age_simple_plot<-age_plot+
  geom_smooth(data=data.frame(age=test_set$age,stroke=y_hat_simple2,gender=test_set$gender),colour="#000
age_simple_plot
```