Project

Predict the insured's risk of heart disease

Zihua Wang

Contents

Introduction:In order to cope with the huge impact that the uncertainty of the insured's potential heart attack has on the insurance company's operations, it is necessary to predict the occurrence and health index of the insured's heart disease. This project will organize and analyze the data provided by insurance companies. Our main idea is to find the relationship between the occurrence of heart disease and other variables in the data set, and use this relationship to predict the risk of heart disease for all other insured people.

1:Define task Objectives and explore data.

The purpose of the project is to obtain a model from the training set data and then use it on testing data set or new data to predict the heart disease risk. So first I would have as the output variable whether a heart attack occurs. Others are temporarily used as input variables, and the model features will be expanded and filtered later. Later I will separate into train data and test data.

```
data_all=read.csv("heart_disease_train_data.csv")
head(data_all)# 22 variables which are Sex,HealthIndex and so on.8000 data in data_all.
```

```
##   Id   State StateCode    Sex HadHeartAttack HealthIndex HeightInMeters
## 1  1 Alabama        AL Female            Yes       99.07           1.65
## 2  2 Alabama        AL Female             No       90.70           1.60
## 3  3 Alabama        AL   Male             No       89.29           1.83
## 4  4 Alabama        AL Female            Yes       99.31           1.50
## 5  5 Alabama        AL Female             No       94.24           1.63
## 6  6 Alabama        AL Female             No       99.44           1.52
##   WeightInKilograms SmokerStatus AlcoholDrinkers HadStroke PhysicalHealthDays
## 1             61.69 Never smoked              No        No                  5
## 2             79.38 Never smoked             Yes        No                  0
## 3            108.41 Never smoked              No        No                  0
## 4             47.17 Never smoked              No       Yes                 30
## 5             72.57 Never smoked              No        No                  0
## 6             51.71 Never smoked              No        No                  0
##   MentalHealthDays DifficultyWalking     AgeCategory HadDiabetes
## 1                3               Yes Age 80 or older         Yes
## 2                0                No    Age 55 to 59          No
## 3                0                No    Age 60 to 64          No
## 4               30               Yes Age 80 or older         Yes
## 5                0                No    Age 65 to 69          No
## 6                0                No Age 80 or older          No
##   PhysicalActivities GeneralHealth SleepHours HadAsthma HadKidneyDisease
```

```
## 1                Yes          Good          6        No                No
## 2                Yes          Good          6        No                No
## 3                Yes     Excellent          6        No                No
## 4                 No          Poor         18        No               Yes
## 5                Yes     Very good          8        No                No
## 6                Yes     Excellent          8        No                No
##   HadSkinCancer
## 1            No
## 2            No
## 3           Yes
## 4           Yes
## 5            No
## 6            No
```

2:Data check and Implement Data Cleaning.

I will check the data for missing values and outliers.

```
missing_values <- which(is.na(data_all), arr.ind = TRUE)# Store the missing value index and where is
#the missing value.
head(data_all[missing_values[,1],])# show the data set which contain missing values.
```

```
##        Id    State StateCode    Sex HadHeartAttack HealthIndex HeightInMeters
## 270   270  Arizona        AZ Female           <NA>          NA             NA
## 330   330  Arizona        AZ   Male           <NA>          NA             NA
## 597   597 Arkansas        AR   Male           <NA>          NA             NA
## 1017 1017 Colorado        CO   Male           <NA>          NA             NA
## 1301 1301 Delaware        DE Female           <NA>          NA             NA
## 1533 1533  Florida        FL   Male           <NA>          NA             NA
##      WeightInKilograms SmokerStatus AlcoholDrinkers HadStroke
## 270                 NA         <NA>            <NA>      <NA>
## 330                 NA         <NA>            <NA>      <NA>
## 597                 NA         <NA>            <NA>      <NA>
## 1017                NA         <NA>            <NA>      <NA>
## 1301                NA         <NA>            <NA>      <NA>
## 1533                NA         <NA>            <NA>      <NA>
##      PhysicalHealthDays MentalHealthDays DifficultyWalking AgeCategory
## 270                  NA               NA              <NA>        <NA>
## 330                  NA               NA              <NA>        <NA>
## 597                  NA               NA              <NA>        <NA>
## 1017                 NA               NA              <NA>        <NA>
## 1301                 NA               NA              <NA>        <NA>
## 1533                 NA               NA              <NA>        <NA>
##      HadDiabetes PhysicalActivities GeneralHealth SleepHours HadAsthma
## 270         <NA>               <NA>          <NA>         NA      <NA>
## 330         <NA>               <NA>          <NA>         NA      <NA>
## 597         <NA>               <NA>          <NA>         NA      <NA>
## 1017        <NA>               <NA>          <NA>         NA      <NA>
## 1301        <NA>               <NA>          <NA>         NA      <NA>
## 1533        <NA>               <NA>          <NA>         NA      <NA>
##      HadKidneyDisease HadSkinCancer
## 270              <NA>          <NA>
```

```
## 330              <NA>          <NA>
## 597              <NA>          <NA>
## 1017             <NA>          <NA>
## 1301             <NA>          <NA>
## 1533             <NA>          <NA>
```

I find that 20 data in "missing values" only remain 4 information out of 22. Only 8 data which only lost 1 information.So my idea is remove these 20 data because these data can not provide any help to built model. As for the 8 data, removing them will not effect so much because we have 8000 data in train_data.

```
k=as.vector(missing_values[,1])
data_all=data_all[-k,]# remove useless data from data set
```

Next I will check the data values which is not reasonable.

```
summary(data_all)
```

```
##        Id          State            StateCode            Sex           
##  Min.   :   1   Length:7972        Length:7972        Length:7972       
##  1st Qu.:2002   Class :character   Class :character   Class :character  
##  Median :4002   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :4002                                                           
##  3rd Qu.:6002                                                           
##  Max.   :8000                                                           
##  HadHeartAttack      HealthIndex     HeightInMeters  WeightInKilograms
##  Length:7972        Min.   : 38.50   Min.   :1.000   Min.   : 36.29   
##  Class :character   1st Qu.: 89.58   1st Qu.:1.630   1st Qu.: 68.49   
##  Mode  :character   Median : 93.83   Median :1.700   Median : 81.65   
##                     Mean   : 92.45   Mean   :1.707   Mean   : 84.48   
##                     3rd Qu.: 97.02   3rd Qu.:1.780   3rd Qu.: 96.16   
##                     Max.   :100.00   Max.   :2.160   Max.   :227.25   
##  SmokerStatus       AlcoholDrinkers     HadStroke          PhysicalHealthDays
##  Length:7972        Length:7972        Length:7972        Min.   : 0.000    
##  Class :character   Class :character   Class :character   1st Qu.: 0.000    
##  Mode  :character   Mode  :character   Mode  :character   Median : 0.000    
##                                                           Mean   : 5.715    
##                                                           3rd Qu.: 6.000    
##                                                           Max.   :30.000    
##  MentalHealthDays DifficultyWalking  AgeCategory        HadDiabetes       
##  Min.   : 0.000   Length:7972        Length:7972        Length:7972       
##  1st Qu.: 0.000   Class :character   Class :character   Class :character  
##  Median : 0.000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 4.625                                                           
##  3rd Qu.: 5.000                                                           
##  Max.   :30.000                                                           
##  PhysicalActivities GeneralHealth        SleepHours      HadAsthma        
##  Length:7972        Length:7972        Min.   :-10.000   Length:7972       
##  Class :character   Class :character   1st Qu.:  6.000   Class :character  
##  Mode  :character   Mode  :character   Median :  7.000   Mode  :character  
##                                        Mean   :  7.018                     
##                                        3rd Qu.:  8.000                     
##                                        Max.   : 23.000                     
##  HadKidneyDisease   HadSkinCancer     
```

```
##  Length:7972        Length:7972
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

The value of SleepHours contains some negative values.And the variable Sex contain 4 categories.

```
negative_index<-which(data_all$SleepHours < 0)
data_all=data_all[-negative_index,]# remove unreasonable values
```

And the variable Sex contain 4 categories.

```
unique(data_all$Sex)# 4 type
```

```
## [1] "Female" "Male"   "female" "male"
```

```
data_all$Sex[data_all$Sex=="female"]<-"Female"
data_all$Sex[data_all$Sex=="male"]<-"Male"
unique(data_all$Sex)# check the result
```

```
## [1] "Female" "Male"
```

```
nrow(data_all)# check number of rows 8000-28-3
```

```
## [1] 7969
```

Separate data into train data and test data.

```
set.seed(25)
random_numbers <- sample(1:7969, 6000, replace = FALSE)# separate 6000 for training data, 1969 for test
x_train=data_all[random_numbers,-5]# input variables
y_train=data_all[random_numbers,5]# output variable
```

3:Exploratory data analysis and visualization(not necessary).

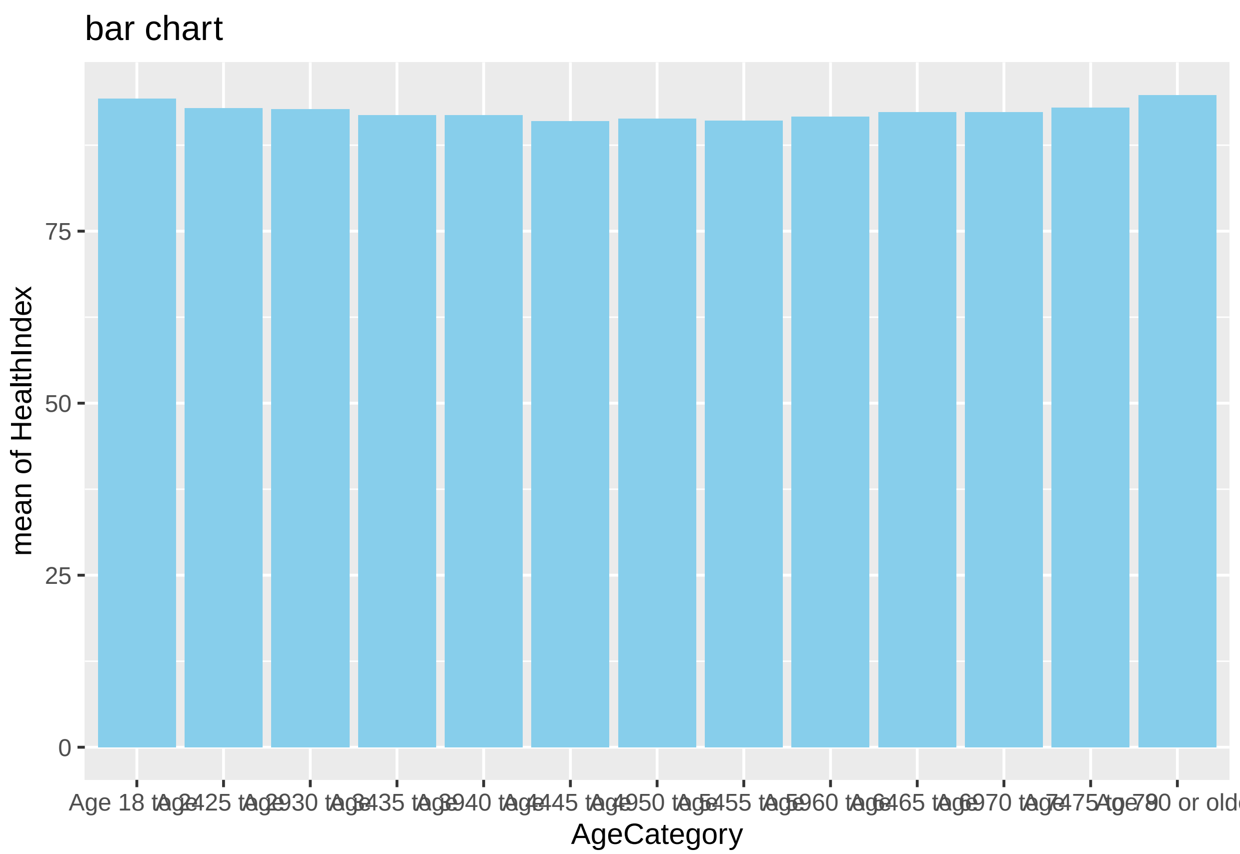
This part focus on exploring relationship not only between output variable and input variables, but also between input variables. And I will visualize with charts and images.

Before I got the model. I am curious about relationship between age and Healthindex.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

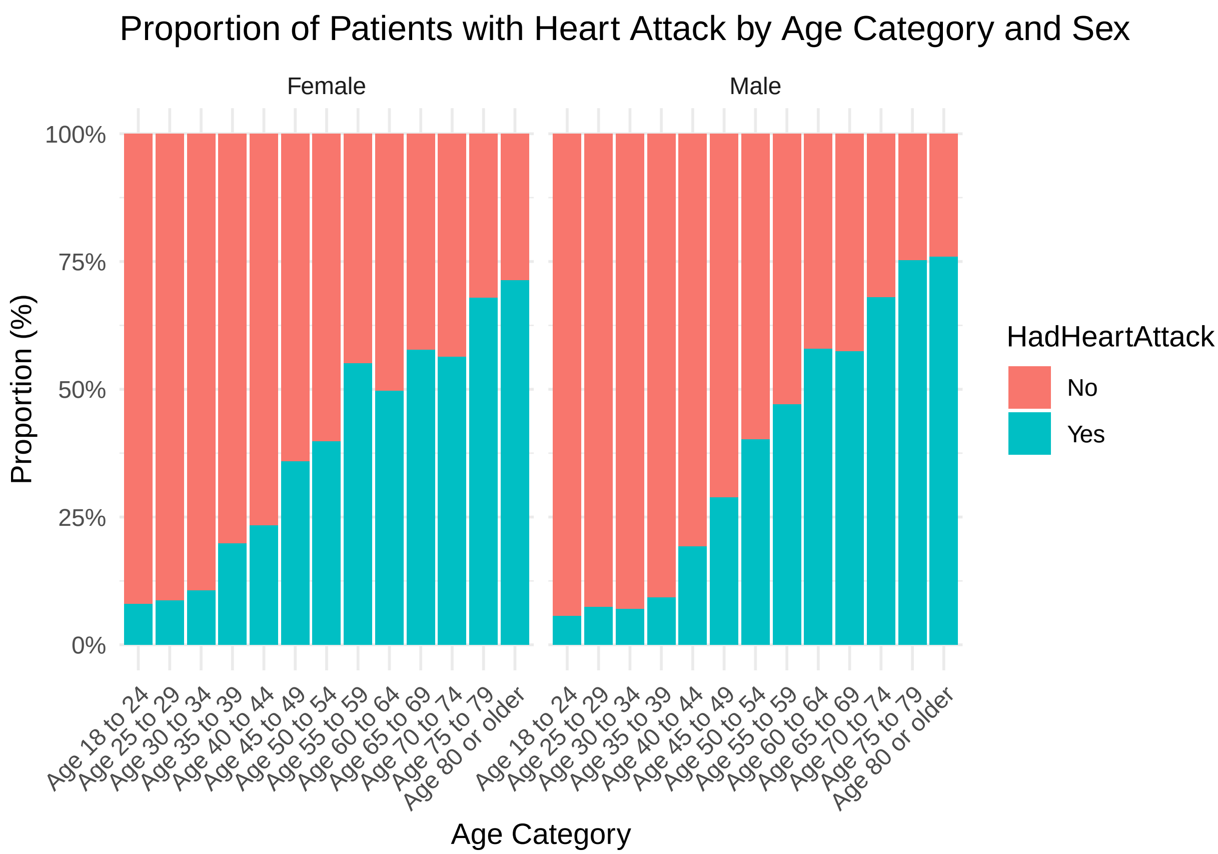
```
ggplot(data_all, aes(x = AgeCategory, y = HealthIndex)) +
  geom_bar(stat = "summary", fun = "mean", fill = "skyblue", position = "dodge") +
  labs(title = "bar chart",
       x = "AgeCategory",
       y = "mean of HealthIndex")
```



Relationship between age, sex and heart attack.

```
ggplot(data_all, aes(x=AgeCategory, fill=HadHeartAttack)) +
  geom_bar(position="fill", aes(y=..prop.., group=HadHeartAttack)) +
  facet_wrap(~Sex) +
  scale_y_continuous(labels=scales::percent) +
  labs(title="Proportion of Patients with Heart Attack by Age Category and Sex",
       x="Age Category",
       y="Proportion (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate X-axis labels for readability
```

```
## Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(prop)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



4:Feature Engineering

In this part I will filter variables.

```r
head(data_all)
```

```
##   Id   State StateCode    Sex HadHeartAttack HealthIndex HeightInMeters
## 1  1 Alabama        AL Female            Yes       99.07           1.65
## 2  2 Alabama        AL Female             No       90.70           1.60
## 3  3 Alabama        AL   Male             No       89.29           1.83
## 4  4 Alabama        AL Female            Yes       99.31           1.50
## 5  5 Alabama        AL Female             No       94.24           1.63
## 6  6 Alabama        AL Female             No       99.44           1.52
##   WeightInKilograms SmokerStatus AlcoholDrinkers HadStroke PhysicalHealthDays
## 1             61.69 Never smoked              No        No                  5
## 2             79.38 Never smoked             Yes        No                  0
## 3            108.41 Never smoked              No        No                  0
## 4             47.17 Never smoked              No       Yes                 30
## 5             72.57 Never smoked              No        No                  0
## 6             51.71 Never smoked              No        No                  0
##   MentalHealthDays DifficultyWalking     AgeCategory HadDiabetes
## 1                3               Yes Age 80 or older         Yes
## 2                0                No    Age 55 to 59          No
## 3                0                No    Age 60 to 64          No
## 4               30               Yes Age 80 or older         Yes
```

```
## 5                 0                No     Age 65 to 69          No
## 6                 0                No Age 80 or older          No
##   PhysicalActivities GeneralHealth SleepHours HadAsthma HadKidneyDisease
## 1                Yes          Good          6        No               No
## 2                Yes          Good          6        No               No
## 3                Yes     Excellent          6        No               No
## 4                 No          Poor         18        No              Yes
## 5                Yes     Very good          8        No               No
## 6                Yes     Excellent          8        No               No
##   HadSkinCancer
## 1            No
## 2            No
## 3           Yes
## 4           Yes
## 5            No
## 6            No
```

```
k=c(1,2,3)
data_all=data_all[,-k]# remove Id, State, StateCode which is useless for predicting result.
x_train=x_train[,-k]
```

4.1 AIC and BIC

Stepwise regression is usually based on linear models. At each step, it considers adding or removing a variable and evaluates the goodness of fit of the model based on predefined criteria (such as AIC, BIC, etc.). Stepwise regression attempts to find an optimal model in a given set of variables, so that the selected model has the minimum information loss or the maximum goodness of fit under the given criteria.

```
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

```
data_transfer<-data_all
data_transfer$HadHeartAttack<-ifelse(data_all$HadHeartAttack=="Yes",1,0)
glm_model=glm(HadHeartAttack~.,data=data_transfer,family=binomial)
AIC_model=step(glm_model,direction = "backward")
```

```
## Start:  AIC=7414.58
## HadHeartAttack ~ Sex + HealthIndex + HeightInMeters + WeightInKilograms +
##     SmokerStatus + AlcoholDrinkers + HadStroke + PhysicalHealthDays +
##     MentalHealthDays + DifficultyWalking + AgeCategory + HadDiabetes +
##     PhysicalActivities + GeneralHealth + SleepHours + HadAsthma +
##     HadKidneyDisease + HadSkinCancer
##
##                      Df Deviance    AIC
## - HealthIndex         1   7340.6 7412.6
## - WeightInKilograms   1   7340.6 7412.6
## - HadSkinCancer       1   7340.6 7412.6
## - PhysicalActivities  1   7340.7 7412.7
## - HadAsthma           1   7341.1 7413.1
```

```
## - SleepHours          1   7341.4 7413.4
## - HeightInMeters      1   7341.7 7413.7
## <none>                    7340.6 7414.6
## - MentalHealthDays    1   7346.1 7418.1
## - AlcoholDrinkers     1   7346.3 7418.3
## - PhysicalHealthDays  1   7346.4 7418.4
## - DifficultyWalking   1   7350.1 7422.1
## - HadKidneyDisease    1   7373.4 7445.4
## - HadDiabetes         3   7408.8 7476.8
## - SmokerStatus        3   7436.6 7504.6
## - HadStroke           1   7470.1 7542.1
## - Sex                 1   7475.0 7547.0
## - GeneralHealth       4   7512.8 7578.8
## - AgeCategory        12   8001.3 8051.3
##
## Step:  AIC=7412.58
## HadHeartAttack ~ Sex + HeightInMeters + WeightInKilograms + SmokerStatus +
##     AlcoholDrinkers + HadStroke + PhysicalHealthDays + MentalHealthDays +
##     DifficultyWalking + AgeCategory + HadDiabetes + PhysicalActivities +
##     GeneralHealth + SleepHours + HadAsthma + HadKidneyDisease +
##     HadSkinCancer
##
##                      Df Deviance    AIC
## - HadSkinCancer       1   7340.6 7410.6
## - WeightInKilograms   1   7340.6 7410.6
## - PhysicalActivities  1   7340.7 7410.7
## - HadAsthma           1   7341.1 7411.1
## - SleepHours          1   7341.4 7411.4
## <none>                    7340.6 7412.6
## - HeightInMeters      1   7343.9 7413.9
## - MentalHealthDays    1   7346.1 7416.1
## - AlcoholDrinkers     1   7346.3 7416.3
## - PhysicalHealthDays  1   7346.4 7416.4
## - DifficultyWalking   1   7350.1 7420.1
## - HadKidneyDisease    1   7373.4 7443.4
## - HadDiabetes         3   7408.9 7474.9
## - SmokerStatus        3   7436.7 7502.7
## - HadStroke           1   7470.1 7540.1
## - Sex                 1   7475.9 7545.9
## - GeneralHealth       4   7513.0 7577.0
## - AgeCategory        12   8005.5 8053.5
##
## Step:  AIC=7410.62
## HadHeartAttack ~ Sex + HeightInMeters + WeightInKilograms + SmokerStatus +
##     AlcoholDrinkers + HadStroke + PhysicalHealthDays + MentalHealthDays +
##     DifficultyWalking + AgeCategory + HadDiabetes + PhysicalActivities +
##     GeneralHealth + SleepHours + HadAsthma + HadKidneyDisease
##
##                      Df Deviance    AIC
## - WeightInKilograms   1   7340.7 7408.7
## - PhysicalActivities  1   7340.8 7408.8
## - HadAsthma           1   7341.2 7409.2
## - SleepHours          1   7341.4 7409.4
## <none>                    7340.6 7410.6
```

```
## - HeightInMeters      1   7344.0 7412.0
## - MentalHealthDays    1   7346.2 7414.2
## - AlcoholDrinkers     1   7346.4 7414.4
## - PhysicalHealthDays  1   7346.4 7414.4
## - DifficultyWalking   1   7350.2 7418.2
## - HadKidneyDisease    1   7373.4 7441.4
## - HadDiabetes         3   7409.0 7473.0
## - SmokerStatus        3   7436.8 7500.8
## - HadStroke           1   7470.1 7538.1
## - Sex                 1   7475.9 7543.9
## - GeneralHealth       4   7513.0 7575.0
## - AgeCategory        12   8027.8 8073.8
##
## Step:  AIC=7408.68
## HadHeartAttack ~ Sex + HeightInMeters + SmokerStatus + AlcoholDrinkers +
##     HadStroke + PhysicalHealthDays + MentalHealthDays + DifficultyWalking +
##     AgeCategory + HadDiabetes + PhysicalActivities + GeneralHealth +
##     SleepHours + HadAsthma + HadKidneyDisease
##
##                      Df Deviance    AIC
## - PhysicalActivities  1   7340.8 7406.8
## - HadAsthma           1   7341.2 7407.2
## - SleepHours          1   7341.5 7407.5
## <none>                    7340.7 7408.7
## - HeightInMeters      1   7344.2 7410.2
## - MentalHealthDays    1   7346.2 7412.2
## - PhysicalHealthDays  1   7346.5 7412.5
## - AlcoholDrinkers     1   7346.5 7412.5
## - DifficultyWalking   1   7350.6 7416.6
## - HadKidneyDisease    1   7373.5 7439.5
## - HadDiabetes         3   7411.2 7473.2
## - SmokerStatus        3   7437.3 7499.3
## - HadStroke           1   7470.2 7536.2
## - Sex                 1   7476.5 7542.5
## - GeneralHealth       4   7514.4 7574.4
## - AgeCategory        12   8047.9 8091.9
##
## Step:  AIC=7406.83
## HadHeartAttack ~ Sex + HeightInMeters + SmokerStatus + AlcoholDrinkers +
##     HadStroke + PhysicalHealthDays + MentalHealthDays + DifficultyWalking +
##     AgeCategory + HadDiabetes + GeneralHealth + SleepHours +
##     HadAsthma + HadKidneyDisease
##
##                      Df Deviance    AIC
## - HadAsthma           1   7341.4 7405.4
## - SleepHours          1   7341.6 7405.6
## <none>                    7340.8 7406.8
## - HeightInMeters      1   7344.4 7408.4
## - MentalHealthDays    1   7346.4 7410.4
## - PhysicalHealthDays  1   7346.8 7410.8
## - AlcoholDrinkers     1   7346.8 7410.8
## - DifficultyWalking   1   7351.3 7415.3
## - HadKidneyDisease    1   7373.6 7437.6
## - HadDiabetes         3   7411.5 7471.5
```

```
## - SmokerStatus        3   7438.4 7498.4
## - HadStroke           1   7470.5 7534.5
## - Sex                 1   7476.6 7540.6
## - GeneralHealth       4   7517.3 7575.3
## - AgeCategory        12   8052.1 8094.1
##
## Step:  AIC=7405.38
## HadHeartAttack ~ Sex + HeightInMeters + SmokerStatus + AlcoholDrinkers +
##     HadStroke + PhysicalHealthDays + MentalHealthDays + DifficultyWalking +
##     AgeCategory + HadDiabetes + GeneralHealth + SleepHours +
##     HadKidneyDisease
##
##                      Df Deviance    AIC
## - SleepHours          1   7342.2 7404.2
## <none>                    7341.4 7405.4
## - HeightInMeters      1   7345.0 7407.0
## - MentalHealthDays    1   7347.1 7409.1
## - AlcoholDrinkers     1   7347.4 7409.4
## - PhysicalHealthDays  1   7347.6 7409.6
## - DifficultyWalking   1   7352.1 7414.1
## - HadKidneyDisease    1   7374.2 7436.2
## - HadDiabetes         3   7412.3 7470.3
## - SmokerStatus        3   7439.0 7497.0
## - HadStroke           1   7471.7 7533.7
## - Sex                 1   7476.6 7538.6
## - GeneralHealth       4   7518.8 7574.8
## - AgeCategory        12   8053.7 8093.7
##
## Step:  AIC=7404.2
## HadHeartAttack ~ Sex + HeightInMeters + SmokerStatus + AlcoholDrinkers +
##     HadStroke + PhysicalHealthDays + MentalHealthDays + DifficultyWalking +
##     AgeCategory + HadDiabetes + GeneralHealth + HadKidneyDisease
##
##                      Df Deviance    AIC
## <none>                    7342.2 7404.2
## - HeightInMeters      1   7345.9 7405.9
## - MentalHealthDays    1   7348.2 7408.2
## - AlcoholDrinkers     1   7348.2 7408.2
## - PhysicalHealthDays  1   7348.4 7408.4
## - DifficultyWalking   1   7352.9 7412.9
## - HadKidneyDisease    1   7375.1 7435.1
## - HadDiabetes         3   7413.6 7469.6
## - SmokerStatus        3   7440.1 7496.1
## - HadStroke           1   7472.7 7532.7
## - Sex                 1   7477.4 7537.4
## - GeneralHealth       4   7520.4 7574.4
## - AgeCategory        12   8062.3 8100.3
```

I choose "HeightInMeters","MentalHealthDays","AlcoholDrinkers","PhysicalHealthDays","DifficultyWalking","HadKidn
by AIC criteria.

4.2 L1 regularization

Lasso regression can be used to process data containing numerical features and categorical features. Lasso regression promotes sparse coefficients of the model by adding L1 regularization, so that feature selection can be performed, that is, the coefficients of some features are penalized to zero. This makes lasso regression very suitable for data sets with a large number of features, including categorical features.

```
j=c(2,3,4,8,9,15)# label the numeric variables
x_train_num=x_train[,j]
x_train_cat=x_train[,-j]
for (i in 1:12){
  x_train_cat[,i]=as.numeric(factor(x_train_cat[,i]))
}
x_train_trans=cbind(x_train_num,x_train_cat)
y_train_trans=as.numeric(factor(y_train))
lasso_model=glmnet(x_train_trans,y_train_trans,alpha=1,lambda = 0.01)
```

```
coef(lasso_model)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)         0.4491226856
## HealthIndex         .
## HeightInMeters      .
## WeightInKilograms   .
## PhysicalHealthDays  0.0053807535
## MentalHealthDays    0.0009843518
## SleepHours          .
## Sex                 0.1337747229
## SmokerStatus       -0.0477893403
## AlcoholDrinkers    -0.0283009359
## HadStroke           0.2080984767
## DifficultyWalking   0.1026823642
## AgeCategory         0.0374411903
## HadDiabetes         0.0594774966
## PhysicalActivities -0.0047709968
## GeneralHealth       .
## HadAsthma           .
## HadKidneyDisease    0.1072957176
## HadSkinCancer       .
```

I choose variables"PhysicalHealthDays","MentalHealthDays","Sex","SmokerStatus","AlcoholDrinkers","HadStroke","Di
by lasso L1 regularization.

4.3 PCA

Principal Component Analysis (PCA) is a commonly used data dimensionality reduction technique used to discover the main patterns in data and reduce the dimensionality of the data. PCA achieves dimensionality reduction and screening of main variables by converting the original data into a new set of linearly independent variables, called principal components.

```
x_train_stand=scale(x_train_trans)# before PCA I will do standardlization for the data set.
PCA_matrix=prcomp(x_train_stand)
summary(PCA_matrix)# I gonna use 90% cumulative proportion.So I choose PC1 to PC13.
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
```

```
## Standard deviation     1.6457 1.4884 1.26560 1.17724 1.01890 0.99696 0.98544
## Proportion of Variance 0.1504 0.1231 0.08899 0.07699 0.05768 0.05522 0.05395
## Cumulative Proportion  0.1504 0.2735 0.36251 0.43951 0.49718 0.55240 0.60635
##                            PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     0.95466 0.93444 0.92442 0.90987 0.8859 0.87695 0.82502
## Proportion of Variance 0.05063 0.04851 0.04747 0.04599 0.0436 0.04272 0.03781
## Cumulative Proportion  0.65698 0.70549 0.75296 0.79896 0.8426 0.88528 0.92309
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.74198 0.69676 0.56646 0.16567
## Proportion of Variance 0.03059 0.02697 0.01783 0.00152
## Cumulative Proportion  0.95368 0.98065 0.99848 1.00000
```

```
u=rep(0,13)
for (i in 1:13){
  u[i]=which.max(abs(PCA_matrix$rotation[,i]))# $rotation is to get loading matrix.
}
choose_name=unique(u)# Using loading matrix to get variables.
```

```
names(x_train_trans)[choose_name]# These variables are selected by PCA.
```

```
##  [1] "DifficultyWalking" "WeightInKilograms" "AgeCategory"
##  [4] "SmokerStatus"      "GeneralHealth"     "SleepHours"
##  [7] "HadKidneyDisease"  "HadStroke"         "AlcoholDrinkers"
## [10] "HadDiabetes"
```

4.4 Conclusion

By comparing the three methods of selecting variables. I found that most of the variables they selected are similar, which also shows that these variables play a decisive role in the establishment of the model to a certain extent.So use variables "HadStroke", "HadKidneyDisease", "AgeCategory", "HadDiabetes", "SmokerStatus","DifficultyWalking","AlcoholDrinkers","PhysicalHealthDays","Sex","MentalHealthDays","GeneralHealth" ,"WeightInKilograms","SleepHours","HeightInMeters","PhysicalActivities". And The last two or three variables are open to question.

5:Model selecting and comparing

In this module I will use logistic regression, MLP, random forest to build models.And retain their prediction set data.

5.1 Logistic regression

```
x_train_stand=as.data.frame(x_train_stand)
l=c(1,16,18)
x_train_real=x_train_stand[-l]# only contain 15 variables.
x_train_matrix=as.matrix(x_train_real)
y_train=y_train_trans-1# make y to be 0 or 1.
data_stand=cbind.data.frame(x_train_real,y_train)# combine x and y
cv_model=cv.glmnet(x_train_matrix,y_train,alpha=0.5,family="binomial",type.measure = "class")# built lo
```

```
x_test=data_all[-random_numbers,]
f=c(2,3,17,19)
x_test=x_test[-f]
g=c(2,3,7,8,14)
```

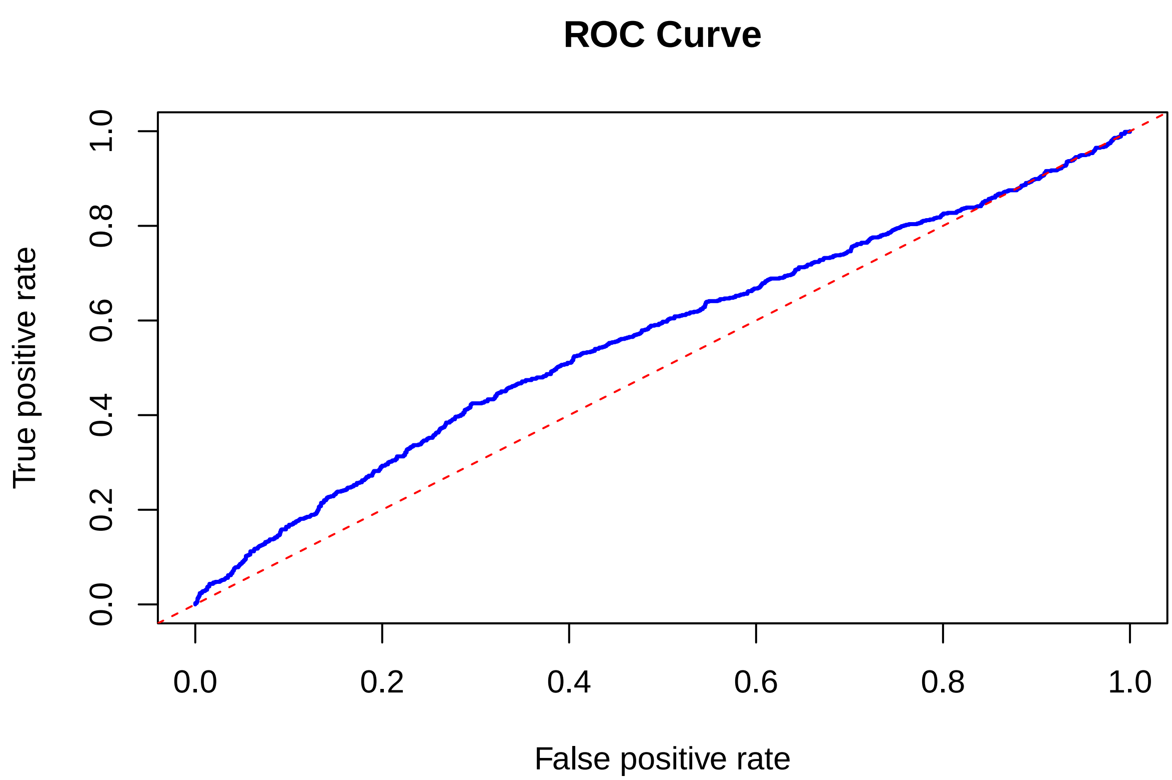
```r
x_test_num=x_test[,g]
x_test_cat=x_test[,-g]
for(i in 1:10){
  x_test_cat[,i]=as.numeric(factor(x_test_cat[,i]))
}
```

```r
x_test=cbind.data.frame(x_test_cat,x_test_num)
x_test=scale(x_test)# get the final standard test data set
result_logis=predict(cv_model,x_test,s=cv_model$lambda.min,type="response")# get the probability
y_test=data_all[-random_numbers,2]
y_test=as.numeric(factor(y_test))-1
```

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.3.2
```

```r
pre<-prediction(result_logis,y_test)
roc<-performance(pre, "tpr", "fpr")
plot(roc, main = "ROC Curve", col = "blue", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "red")
```




```r
predic=ifelse(result_logis>0.9,1,0)
sum((predic-y_test)^2)# test error is too big which means logistic regression is not the suitable model
```

```
## [1] 708
```

5.2 Random Forest

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.2

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
data_all_6=data_all[random_numbers,]
data_all_6$HadHeartAttack=factor(data_all_6$HadHeartAttack)
random_model<-randomForest(HadHeartAttack ~ ., data = data_all_6, ntree = 100)
```

```r
data_test=data_all[-random_numbers,]
data_test_x=data_test[-2]
data_test_y=data_test[2]
pre_rand=predict(random_model,data_test_x)
i=(as.numeric(factor(pre_rand)))-1
sum((i-y_test)^2)# test error is almost 0.25
```

```
## [1] 485
```

```r
pre_rand1=predict(random_model,x_train)
u=as.numeric(factor(pre_rand1))-1
sum((u-y_train)^2)# train error=1/300 which means the model overfit
```

```
## [1] 19
```

```r
random_model<-randomForest(HadHeartAttack ~ ., data = data_all_6, ntree = 100,max.features=10)
pre_rand=predict(random_model,data_test_x)
i=(as.numeric(factor(pre_rand)))-1
sum((i-y_test)^2)# test error is still almost 0.25
```

```
## [1] 496
```

5.3 MLP

```r
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.3.2
```

```r
set.seed(15)
mlp_model <- nnet(HadHeartAttack~., data = data_all_6, size = 5, maxit = 1000)
```

```
## # weights:  191
## initial  value 3904.949532
## iter  10 value 3610.484300
## iter  20 value 3264.015029
## iter  30 value 2942.460042
## iter  40 value 2807.643910
## iter  50 value 2722.711650
## iter  60 value 2704.662732
## iter  70 value 2693.260275
## iter  80 value 2689.350034
## iter  90 value 2680.812344
## iter 100 value 2674.725342
## iter 110 value 2668.058974
## iter 120 value 2663.549488
## iter 130 value 2660.125611
## iter 140 value 2658.190697
## iter 150 value 2657.756282
## iter 160 value 2657.098358
## iter 170 value 2656.356567
## iter 180 value 2655.600254
## iter 190 value 2654.552828
## iter 200 value 2653.265027
## iter 210 value 2651.795420
## iter 220 value 2651.507976
## iter 230 value 2651.494558
## final  value 2651.494502
## converged
```

```r
pre_mlp <- predict(mlp_model, newdata = data_test_x, type = "class")
h=(as.numeric(factor(pre_mlp))-1)-y_test
sum(h^2)# test error also almost 0.25.
```

```
## [1] 476
```

6:Conclusion

I used three models, among which logistic regression has the worst accuracy, while the errors of mlp and random forest are basically around 0.25, which shows that the model still has overfitting, which may be due to too many feature selections or the existence of the data set Some uncertain relationships.