King Abdullah University of Science and

Technology, CEMSE, STAT

Analyzing Heart Attack Possibility using Logistic Regression

By:

Mohammed Al Saleem 188301

Abstract

Heart attacks are dangerous situations, so it is important to find a way to predict them and know the people who have high chances to get heart attacks. High chance or low chance is a classification problem, and Logistic Regression is a classification algorithm that is used to predict a categorical variable. For analysis in this report, heart disease dataset from UCI Machine Learning Repository will be used. From 13 attributes the individual will be classified to have high chance of heart attack or low chance, attributes like age, sex, etc. Moreover, backward stepwise selection is done to find the most significant predictors.

Table of Contents

Table of Figures

INTRODUCTION

The dataset is about the heart attacks (Myocardial Infarction). A heart attack happens when the heart muscle starts to die due to the blood flow to the heart is reduced or stopped. Where the arteries (coronary) that supply blood to heart blocked by buildup of cholesterol, fat, and others.
Some symptoms of heart attacks are chest pain (angina), breath shortness, cold sweet, light head, and nausea. Moreover, heart attacks can hurt or destroy part of the heart muscle.

Scientists collected data of selected factors where they believe these factors have the most effect on the chance of heart attacks. Due to the importance of taking care of heart attacks and its negative consequence, having a measurement which can classify the individuals becomes a necessity. Taking into consideration the main factors that can affect the heart attacks we can have a decision about the situation of the individuals.

Logistic regression is an appropriate regression analysis to conduct and to categorize the heart attacks, where the dependent variable is dichotomous (binary). Logistic regression analyses are a predictive analysis so we can use it to predict after building the model. In addition, it is used to describe the data and to explain the relationship between the dependent variables and independent variables to finds out what are the major attributes that cause heart attacks, and to avoid heart attacks as possible.

Data Description

For analysis in heart attacks, the heart disease dataset from UCI Machine Learning Repository will be used in this report. The objective of this dataset is to predict the individual's situation based on diagnostic measurements to know whether a patient has chance to get heart attacks or no. The original database contains 76 attributes, but only 14 of them will be used same as the most published experiments. The "target" attribute refers to the heart disease presence. 1 represent more chance of heart attack, and 0 no/less chance of heart attack.

There are 303 datapoints, each datapoint represents a patient with his/her measurements of the independent variables and their chance of heart attack.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0	2	0	3	1
9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Figure 1. The head points in the dataset

Attributes Information:

1. (**age**) in years
2. (**sex**) (1 = male; 0 = female)
3. (**cp**) chest pain type (1= typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4. (**trestbps**) resting blood pressure (in mm Hg)
5. (**chol**) serum cholestoral in mg/dl
6. (**fbs**) fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. (**restecg**) resting electrocardiographic results
(0= normal, 1=having ST-T wave abnormality, 2= showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. (**thalach**) maximum heart rate achieved
9. (**exang**) exercise induced angina (1 = yes; 0 = no)
10. (**oldpeak**) ST depression induced by exercise relative to rest
11. (**slope**) the peak exercise ST segment slope (1= upsloping, 2= flat, 3= down sloping)
12. (**ca**) number of major vessels (0-3) colored by fluoroscopy
13. (**thal**) (1 = normal; 2 = fixed defect; 3 = reversable defect)
14. (**target**) the predicted attribute (0= less chance of heart attack, 1= more chance of heart attack)

STATISTICAL METHODS

As shown in the figure 1, the data has 13 independent variables and a binary response variable (y). So, we need a model that captures as much information as possible of the response variable. However, because we have the response variable y with only two possible outputs (0 and 1), the most appropriate model to the data is the logistic regression.

Logistic regression model can be defined as:

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0+\boldsymbol{\beta_1 x}}}{1+e^{\beta_0+\boldsymbol{\beta_1 x}}}.$$

Also, it can be written as:

$$g\big(\pi(\boldsymbol{x})\big) = logit\big(\pi(\boldsymbol{x})\big) = log(odds) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{\beta_1 x}.$$

This transformation is called the logit transformation or log-odds. Also, the main difference between linear regression and logistic regression is that the expected value of the response variable y given $\boldsymbol{x}$ will be between 0 and 1: $0 \leq E(y|\boldsymbol{x}) \leq 1$, while the expected value of the response variable y given $\boldsymbol{x}$ of the linear regression can be any value (-∞, ∞). Furthermore, in the simple linear regression (i.e., $y = \beta_0 + \boldsymbol{\beta_1 x}$) one unit increase in $\boldsymbol{x}$ gives a change in y by $\boldsymbol{\beta_1}$ units. But in the logistic regression, changing one unit in $\boldsymbol{x}$ will give a change in the log-odds by $\boldsymbol{\beta_1}$, not changing the value of y directly. Furthermore, the error term in logistic regression is not normally distributed as in linear regression. it is binomially distributed with mean 0 and variance of $\pi(\boldsymbol{x}) * (1 - \pi(\boldsymbol{x}))$. $\varepsilon \sim bin(0, \pi(\boldsymbol{x}) * \big(1 - \pi(\boldsymbol{x})\big)$.

After fitting the model with all variables (glm(target ~ ., data = heart, family= binomial)), backward stepwise selection will be used to select the final model that have all variables significant.

To measure the goodness of fit, all of PRESS, AIC, ROC, and Confusion Matrix are used. PRESS (Predicted Residual Error Sum of Squares) is a cross-validation form with equation:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

Where $\hat{y}_{i,-i}$ equal to the predicted value of fitting the model after omitting observation i.

AIC (Akaike Information Criteria), it is an analogous to adjusted R^2 and is a measure of fit, which penalizes the model for the number of independent variables, where the model with a minimum AIC value is preferred.

$$AIC = -2l + 2(p + 1),$$

where l is the maximized log likelihood of the model.

ROC (Receiver Operating Characteristic), the ROC curve is a plot with y-axis of the sensitivity values and against x-axis with 1-specificity values, and evaluated at different threshold values c from 0 to 1.

AUC is a performance measure of a model across all possible classification thresholds. AUC equal the area under the ROC curve, where 1 represent the perfect classifier, and 0.5 represent the random classifier. So, the model with higher AUC has better performance of classification between the positive and negative classes.

Confusion matrix is a tabular representation of observed target vs predicted target and it is used to quantify the accuracy of the model. Sensitivity is the proportion of correct identified of positives. Specificity is the proportion of correct identified of negatives.

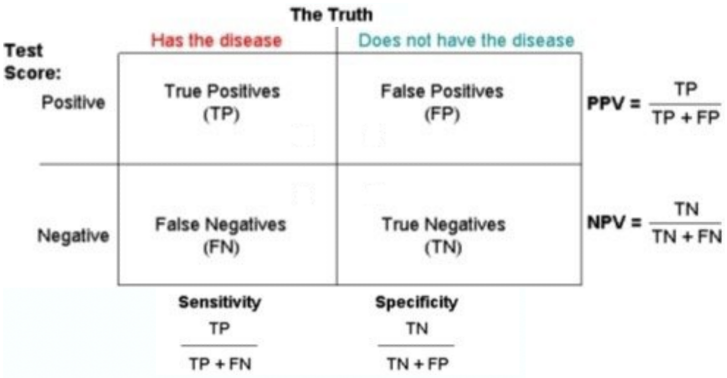


Figure 2. The Confusion matrix

A probability threshold c used to classify individuals as $\hat{Y}$ =1 if $\hat{\pi} >$ c and $\hat{Y}$ = 0 if $\hat{\pi} \leq$ c.

APER = (FN + FP)/(TN+TP+FN+FP)

Accuracy = 1-APER = (TN + TP)/(TN+TP+FN+FP)

APER (Apparent Error Rate) is the fraction of observations that are misclassified by the classification function.

ANALYSIS AND RESULTS

By illustrating the target against the variables as boxplots as figure 3, we can see there are no differences in target with age, trestbps, chol, fbs, restecg, and slope.

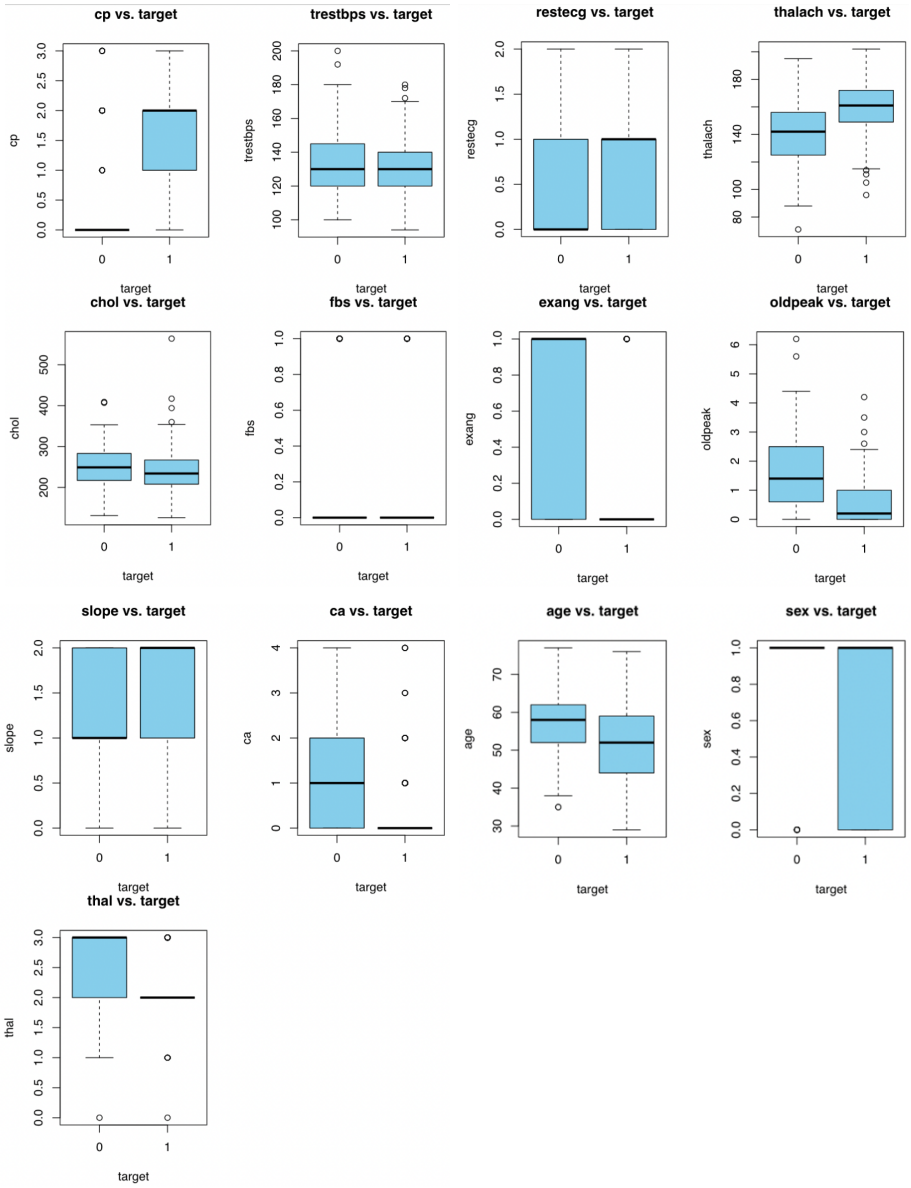


Figure 3. The boxplots of data

Fitting model with all the 13 variable (glm(formula = target ~ ., family = binomial) returns multiple variables not significant. AIC = 239.44, PRESS = 243.4828

The backward stepwise selection applied by deleting the variable of the biggest not significant (a=0.05) p-values one by one, until we end with model7. AIC = 237.41, PRESS = 238.6847

model7 variables are all significant, and it have better results of AIC and PRESS compared with model0.

From the ROC plot curve and the high value of AUC (0.911), we conclude that the model7 have a good fit.

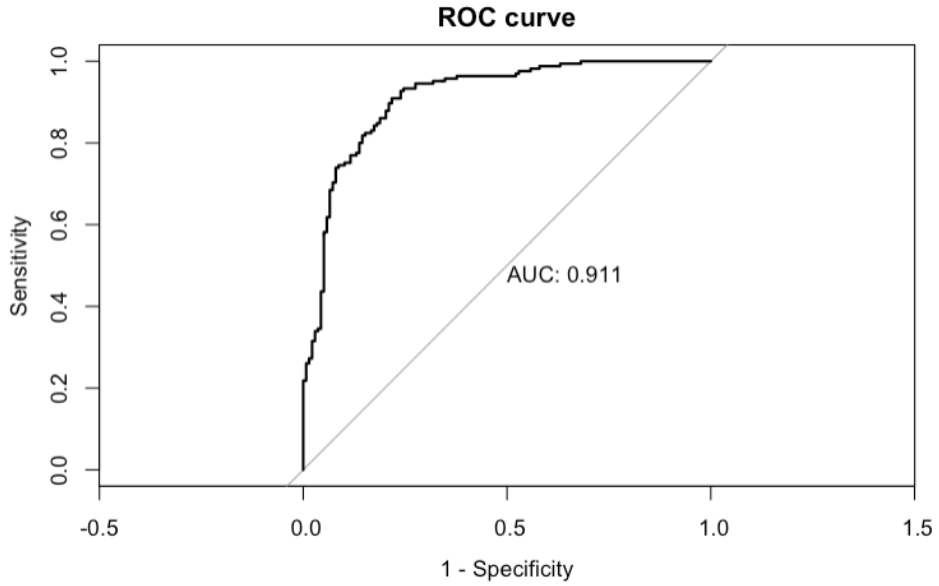


Figure 4. The ROC plot and AUC

The final model (model7): $log(odds) = -1.3898 \times sex + 0.7861 \times cp + 0.0261 \times thalach - 1.0130 \times exang - 0.7262 \times oldpeak - 0.7053 \times ca - 0.8674 \times thal$

Confusion matrix with probability threshold c = 0.5:

model7.pred	0	1
azero	107	15
one	31	150

Accuracy = 1-APER = 0.84488
Sensitivity = 0.77536
Specificity = 0.90909

Figure 5. Model7 confusion matrix results.

DISCUSSION

This report went through analyzing the heart disease dataset to come out with a model that results in the highest level of accuracy to determine whether the person have high chance of heart attack or not. The first step was introducing and understanding the data in hand. The data consisted of 13 independent variables (age, sex, cp, …etc.) and one dependent variable y (more chance of heart attack = 1, less chance of heart attack = 0). After that, the model selection process started to choose the most appropriate model for further analysis.

Since the response variable in the data has only 2 values, it is obvious that logistic regression is the most appropriate model in our case. To find the model with the most significant variables, backward stepwise selection is the most suitable to delete the non-significant variables.

From the confusion matrix, all the accuracy, sensitivity, and specificity have big value which tell the predict is good. But to be sure about the probability threshold value, we test probability thresholds values and search for maximum accuracy, and maximum sensitivity × specificity.

The tested values of probability threshold were 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53. After comparing we conclude that 0.5 probability threshold have the greater accuracy (0.8482), and the greater combination of sensitivity and specificity (0.7049).

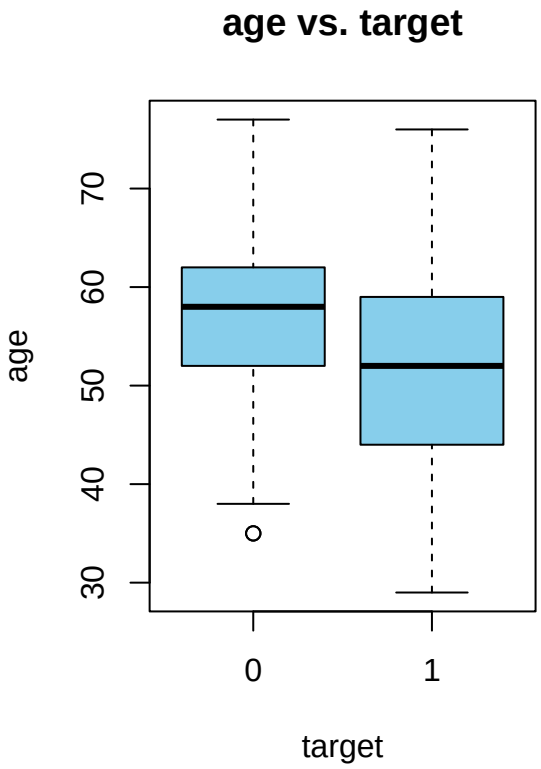
CONCLUSIONS AND RECOMMENDATIONS

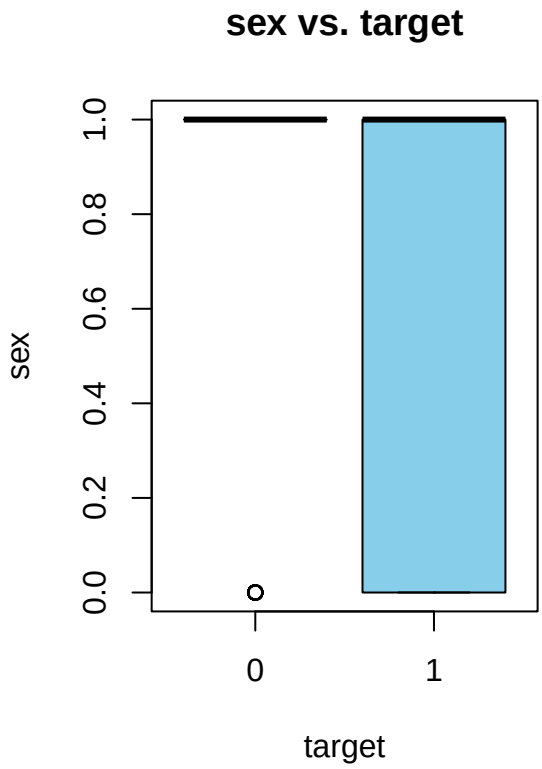
The logistic regression used to fit the model and achieving our goals of understanding the dataset and the factors that related to the heart attacks, and finding out what are the major attributes that cause heart attacks to find the final model. This research and model will help the medical sector to diagnosing if the person expected to get a heart attack, and to avoid heart attacks as possible, where the model will help us understand the causes of heart attacks.

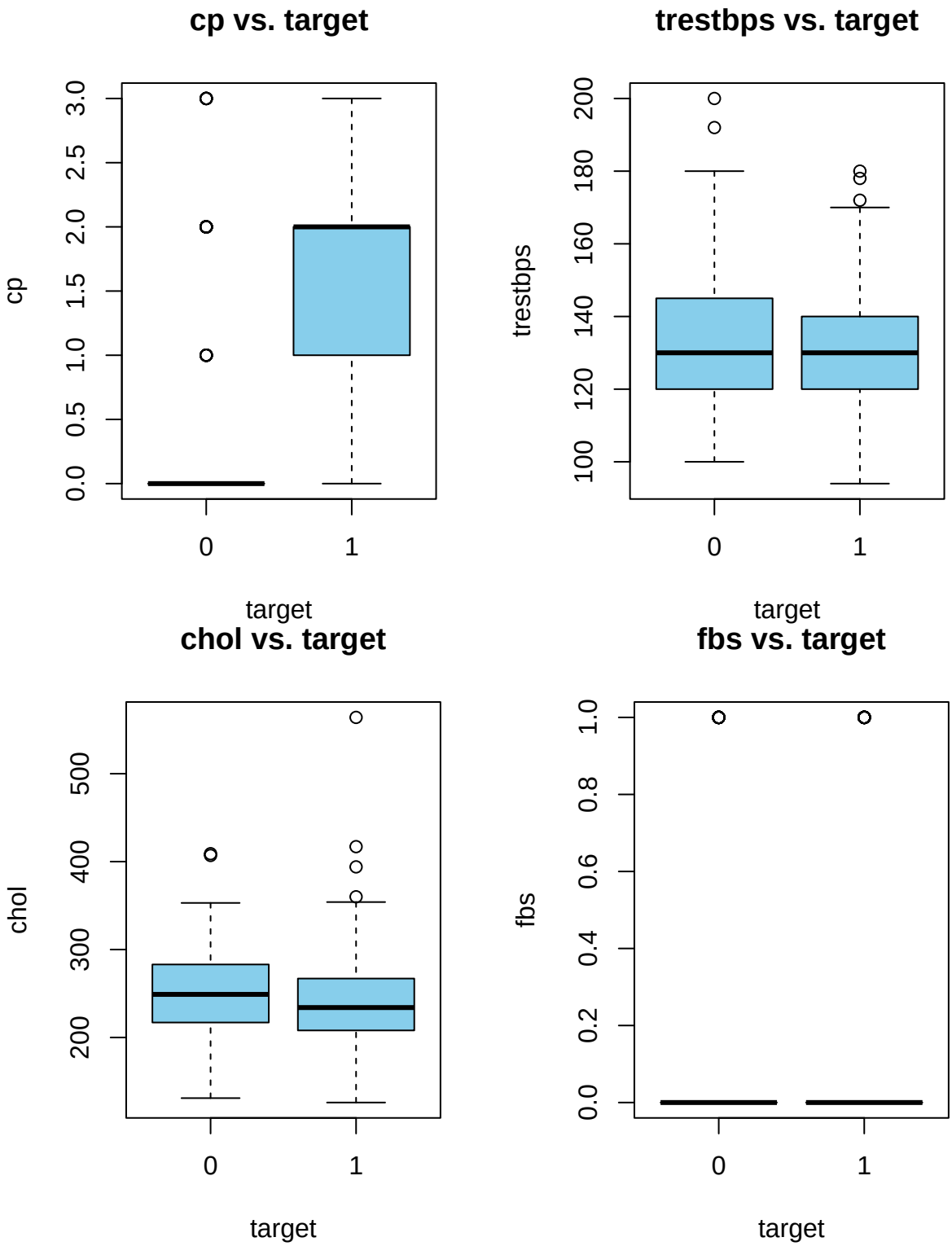
The model could be improved by analyzing the original database that contains the 76 attributes, or other attributes that did not recorded. Also, it is possible to use bootstrap and simulation to get bigger dataset. Furthermore, this methodology can be used to predict other diseases possibilities and avoid them and their complications.

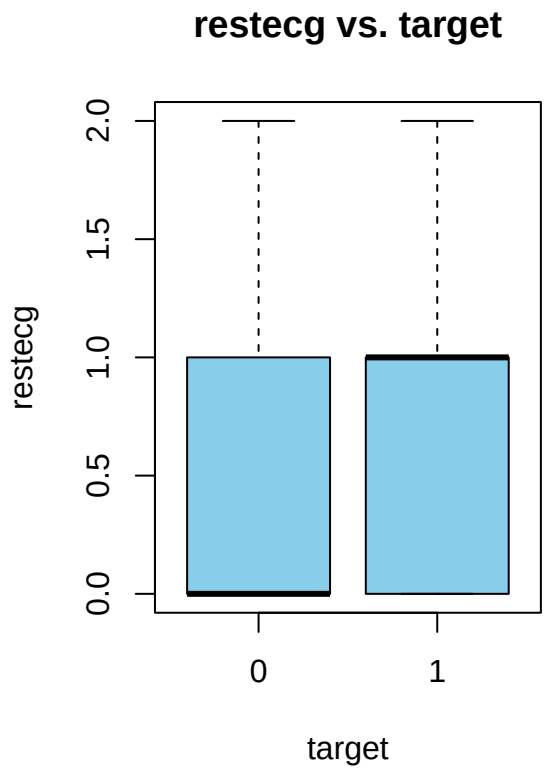
APPENDIX

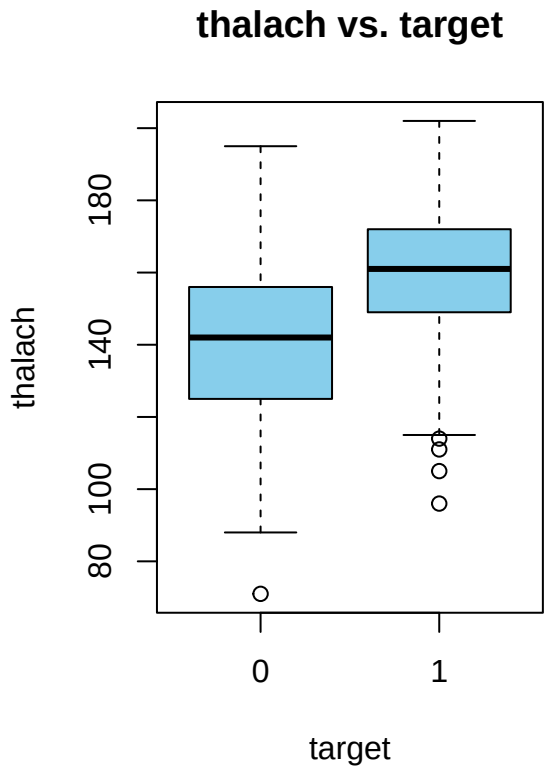
```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

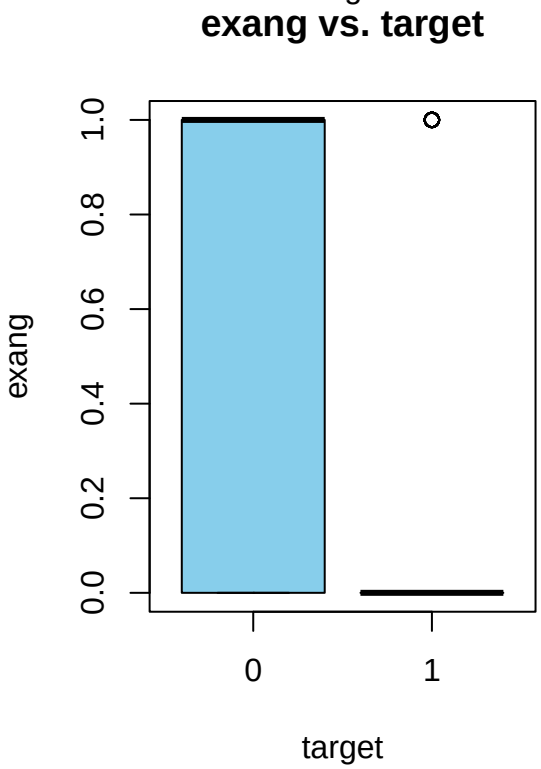


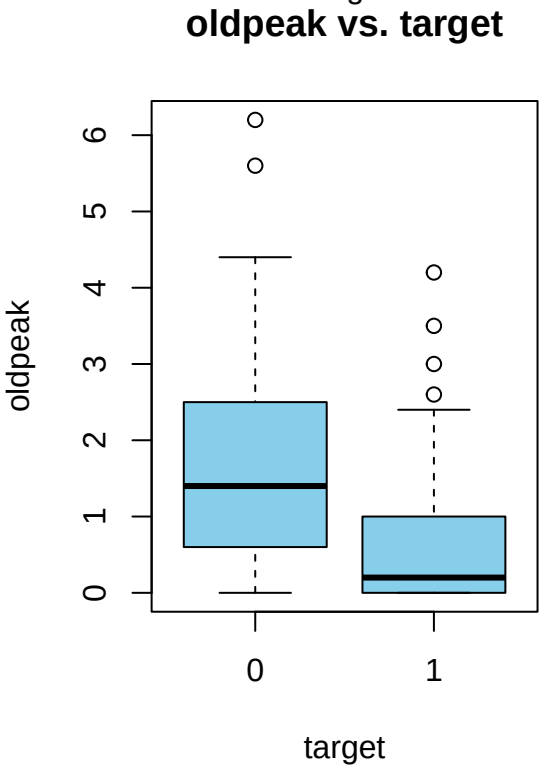


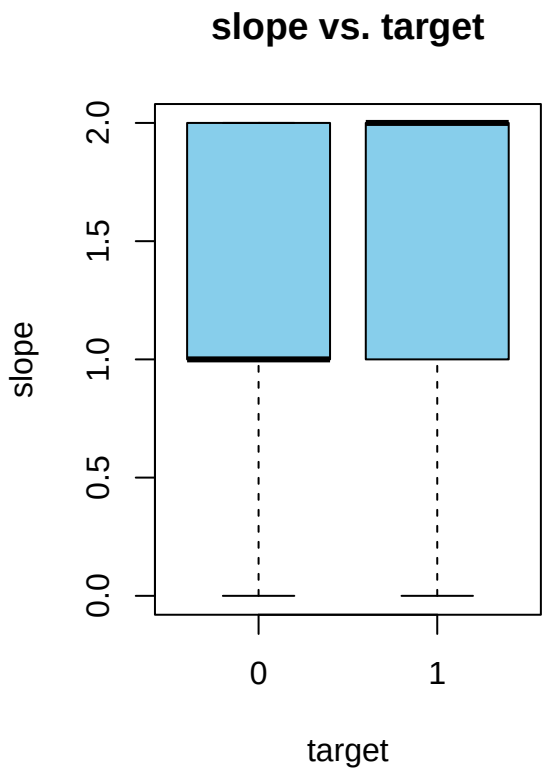

cp vs. target
cp
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
target
trestbps vs. target
trestbps
100
120
140
160
180
200
0
1
target
chol vs. target
chol
200
300
400
500
0
1
target
fbs vs. target
fbs
0.0
0.2
0.4
0.6
0.8
1.0
0
1
target

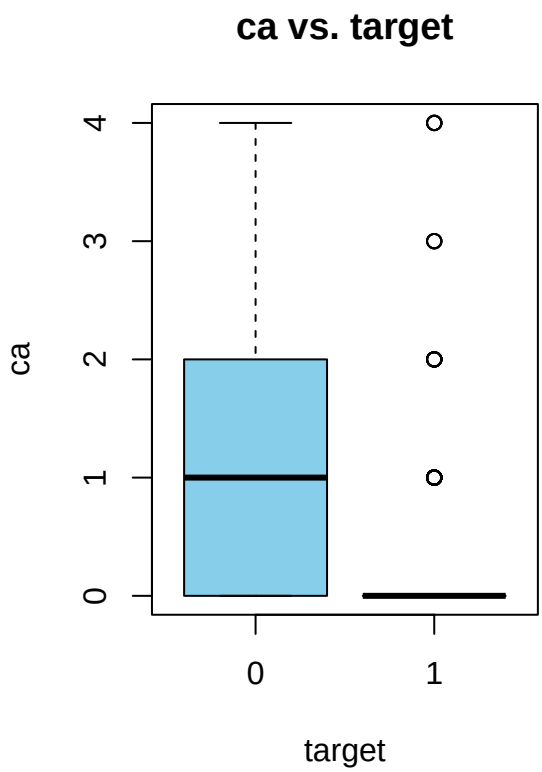

restecg vs. target
restecg
0.0
0.5
1.0
1.5
2.0
0
1
target

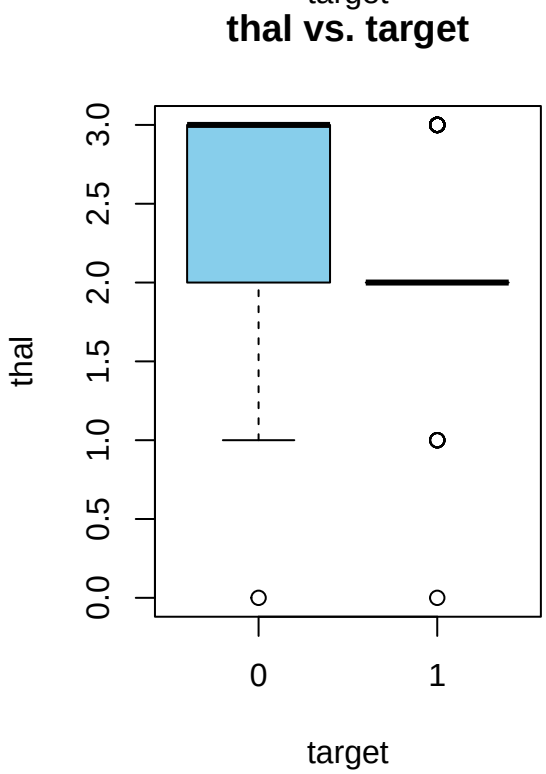

thalach vs. target
thalach
80
100
140
180
0
1
target


exang vs. target
exang
0.0
0.2
0.4
0.6
0.8
1.0
0
1
target


oldpeak vs. target
oldpeak
0
1
2
3
4
5
6
0
1
target







from the Box plots we can see there are no differences in targets with age, trestbps, chol, fbs, restecg, and slope.

```
## $call
## glm(formula = target ~ ., family = binomial, data = heart)
##
## $coefficients
##                 Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)  3.450472415 2.571479409  1.34182386 1.796531e-01
## age         -0.004908470 0.023175419 -0.21179638 8.322659e-01
## sex         -1.758180738 0.468774067 -3.75059301 1.764168e-04
## cp           0.859850938 0.185397070  4.63788848 3.519866e-06
```

```
## trestbps    -0.019476620 0.010338612 -1.88387189 5.958231e-02
## chol        -0.004630231 0.003782218 -1.22421076 2.208727e-01
## fbs          0.034887645 0.529465144  0.06589224 9.474636e-01
## restecg      0.466282248 0.348269280  1.33885552 1.806177e-01
## thalach      0.023210935 0.010459953  2.21902868 2.648477e-02
## exang       -0.979980686 0.409784181 -2.39145563 1.678171e-02
## oldpeak     -0.540273946 0.213849080 -2.52642633 1.152296e-02
## slope        0.579288142 0.349806611  1.65602400 9.771696e-02
## ca          -0.773349275 0.190884846 -4.05139167 5.091390e-05
## thal        -0.900431861 0.290098271 -3.10388566 1.909971e-03
##
## $aic
## [1] 239.436
```

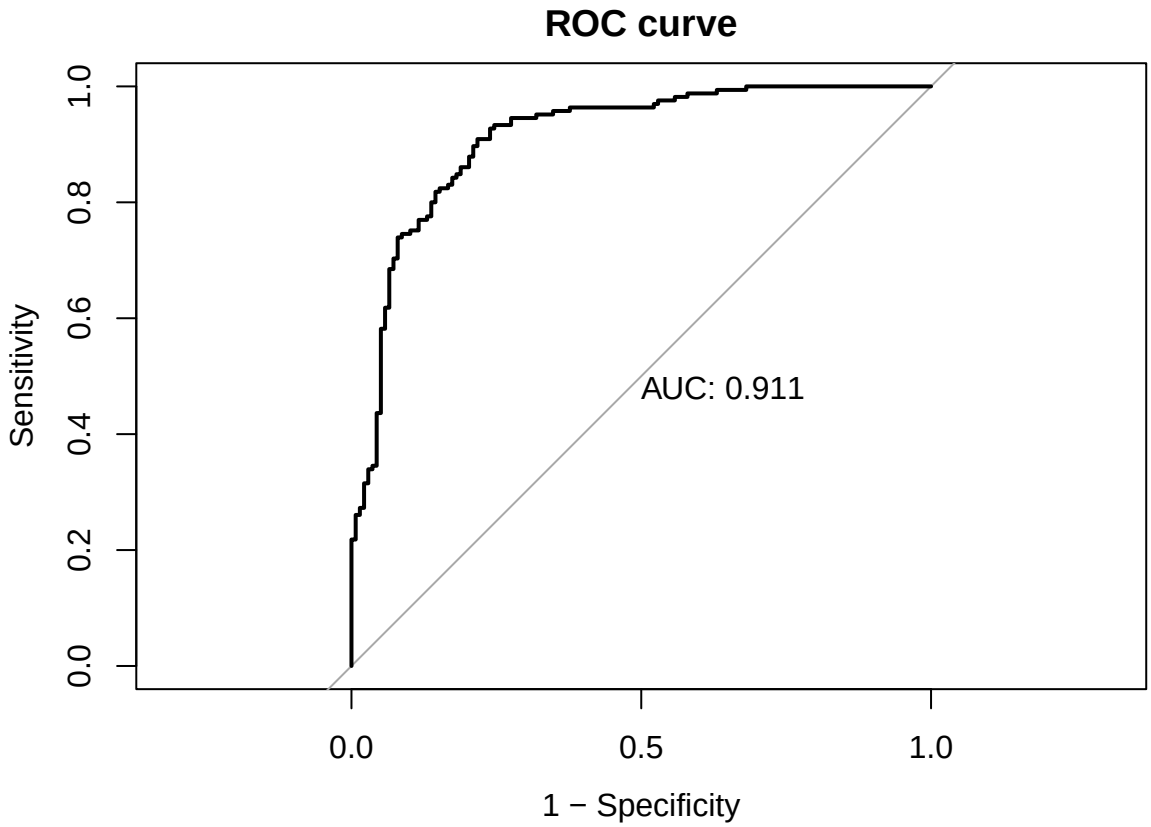
```
## [1] "PRESS = 243.4828"
```

(Backward stepwise selection) After deleting biggest not significant (a=0.05) p-values one by one, we end with model7.

```
## $call
## glm(formula = target ~ . - fbs - age - chol + 0 - trestbps -
##     slope - restecg, family = binomial, data = heart)
##
## $coefficients
##            Estimate  Std. Error   z value     Pr(>|z|)
## sex     -1.38979751 0.405034873 -3.431303 6.006885e-04
## cp       0.78607443 0.174309928  4.509637 6.493878e-06
## thalach  0.02606537 0.004413455  5.905887 3.507536e-09
## exang   -1.01301329 0.375810030 -2.695546 7.027334e-03
## oldpeak -0.72617158 0.175797166 -4.130735 3.616046e-05
## ca      -0.70527509 0.172538124 -4.087648 4.357690e-05
## thal    -0.86739170 0.258816723 -3.351374 8.041155e-04
##
## $aic
## [1] 237.4102
```

```
## [1] "PRESS = 238.6847"
```

model7 variables are all significant, and it have better results of AIC and PRESS compared with model0. AIC decreased from 239.44 to 237.41, and PRESS decreased from 243.4828 to 238.6847.



From the ROC plot curve and the high value of AUC equal to 0.911, we conclude that the model7 have a good fit.

The final model: log(odds) = -1.3898×sex + 0.7861×cp + 0.0261×thalach - 1.0130×exang - 0.7262×oldpeak - 0.7053×ca - 0.8674×thal

```
##
## model7.pred   0   1
##       azero 107  15
##       one    31 150
```

```
## [1] "Accuracy = 0.8482"
```

```
## [1] "Sensitivity = 0.7754"
```

```
## [1] "Specificity = 0.9091"
```

all Accuracy, Sensitivity, and Specificity have big value which tell the predict is good.

```
## [1] "For cut 0.3: Accuracy = 0.8251, Sensitivity * Specificity = 0.644"
## [1] "For cut 0.4: Accuracy = 0.8482, Sensitivity * Specificity = 0.6966"
## [1] "For cut 0.5: Accuracy = 0.8482, Sensitivity * Specificity = 0.7049"
## [1] "For cut 0.6: Accuracy = 0.8284, Sensitivity * Specificity = 0.6859"
## [1] "For cut 0.7: Accuracy = 0.8185, Sensitivity * Specificity = 0.6751"
```

We can see that cut = 0.5 have the greater accuracy (0.8482), and the greater combination of sensitivity and specificity (0.7049).

DATA SOURCE:

Heart Disease Data Set. UCI Machine Learning Repository: Heart disease data set. (n.d.). Retrieved November 13, 2022, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.