Title: Analyzing the Presence of Heart Attack

Author: Mei Ying Tan

Abstract: In this paper, a dataset with binary response was used for analysis and the dataset was split randomly into training and testing sets and prediction performance was evaluated on the test set. For this dataset, conventional logistic regression was applied for predicting the presence of heart attack based on the observed subjects' characteristics. Logistic regression with LASSO penalty and elastic net were also fitted and overall, conventional logistic regression yielded the best final model since it gave the least number of covariates and its correct classification rate as well as the AUC of the ROC were comparable to the others.

1.0 - INTRODUCTION

Coronary Heart Disease (CHD) is a heart disease where it is caused by the accumulation of fatty materials in the blood vessels that supply the heart with oxygen and resulting in a heart attack, chest pain or angina. According to the Centers for Disease Control and Prevention (CDC), about 655,000 Americans die from heart disease each year - approximately 1 in every 4 deaths. Given the fact that heart disease is one of the leading causes of death among men and women in the United States, some questions about the risk factors of heart disease have naturally risen. Our group was motivated to identify the risk factors that would result in a higher or more chance of getting a heart disease, specifically heart attack. The data set was found online from the UCI website: https://archive.ics.uci.edu/ml/datasets/Heart+Disease and it was originally retreived from a Cleveland database. The data set used had a sample size of 303 and consisted of a total number of 12 attributes that were potentially contributing to the certain cardiovascular events or the presence of heart attack.

1.1 - METHODS

Upon fitting the models, we first did some data modification. We divided the dataset into two subsets where the first subset called the training dataset was used to fit the model while the second dataset called the test dataset was used to validate or test the model built or in other words, to evaluate the model performance. In this analysis, we allocated 70% of the dataset into the training dataset randomly while the other 30% was allocated into the test dataset. For the heart disease dataset in part 1, since the response variable was binary, we first used the training dataset to fit the model using the conventional logistic regression with the logit link function and then we conducted a model selection process, particularly stepwise selection procedure to only select a reduced number of significant predictors for model building. However, due to the limitations and criticisms of stepwise selection procedure, we had also fitted the Logistic Regression with Lasso Shrinkage Method into our model since Lasso Shrinkage Method had the ability to conduct model fitting and model selection simultaneously as well as imposing one or more model restriction specifically L1 penalty which resulted in reducing the prediction error variance when dealing with high-dimensional data. Aside from that, we also fitted Elastic Net into our model due to the fact that Lasso would not work well in model fitting if the pairwise correlations between the predictors were very high.

Finally, we evaluated the model performances among the three methods: Conventional Logistic Regression, Logistic Regression with Lasso penalty, and Elastic Net and chose the best final model via the performance of the test dataset in terms of its correct classification rate and area under the curve of the Receiver Operating Characteristic (ROC) curve.

1.2 - EXPLORATORY DATA ANALYSIS

The presence of heart disease dataset had a sample size of 303 and no missing values were found. After performing the data cleaning process, we found a duplicate observation and proceeded to remove that observation as they were not helpful in adding additional information into the model. Thus, our updated sample size was 302. The response variable consisted of a binary outcome where “0” represents no or less chance of heart attack, while “1” represents more chances of getting a heart attack. Furthermore, the dataset used was a balanced data since the response variable for the event of interest (more chance of getting a heart attack) constituted about 54.30% of the total sample size. A total of 12 predictors were used in the analysis and the detailed information of the respective variables was shown in Table 1.

Variable	**Type**	**Levels**	**Description**
Age	Numerical		Age of the subjects
Sex	Categorical	0 = Female; 1 = Male	Sex of the subjects

Cp	Categorical	0 = Typical angina;
		1 = Atypical angina;
		2 = Non-anginal pain;
		3 = Asymptomatic
			Chest pain type

Trestbps	Numerical		Resting blood pressure (in mm Hg on admission to the hospital)
Chol	Numerical		Serum cholesterol in mg/dl
Fbs	Categorical	0 = False; 1 = True	Fasting blood sugar > 120 mg/dl
Restecg	Categorical	0 = Normal; 1 = Having ST-T wave abnormality T wave inversions and/or ST elevation or depression of > 0.05 mV; 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria	Resting electrocardiographic results
Thalach	Numerical		Maximum heart rate achieved
Exang	Categorical	0 = No; 1 = Yes	Exercise induced angina
Oldpeak	Numerical		ST depression induced by exercise relative to rest

Slope	Categorical	0 = Upsloping;
		1 = Flat;
		2 = Downsloping
			Slope of the peak exercise ST segment

Ca	Numerical		Number of major vessels (CA) colored by fluoroscopy

***Table 1:** This table includes all the variables used in this study. It includes the variable name, variable type (numerical or categorical), variable levels, and variable description.*

We had also examined the correlation matrix of the variables. Since all of the correlation coefficients did not exceed the absolute value of 0.7, hence there was no indication of multicollinearity. In particular, there was a negative correlation between ST depression induced by exercise relative to rest and the slope of the peak exercise ST segment with a correlation coefficient of -0.58. Conversely, there was a positive correlation between chest pain type and the presence of heart attack with a correlation coefficient of 0.43.

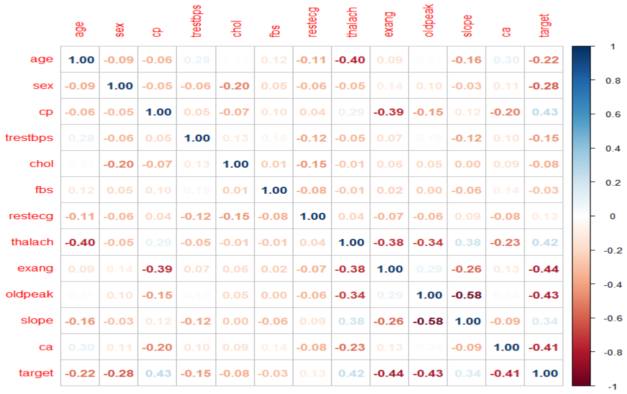


Figure 1: *Correlation matrix of the 12 predictors as well as the response variable in the presence of heart disease dataset.*

To get an overview of the predictors, we plotted density plots of the response variable versus each of the 12 predictors as shown in figure 2. From the density plots, we were able to observe that there were more chances of getting a heart attack if there was no exercise induced angina; if the ST depression induced by exercise relative to rest was low; if the slope of the peak exercise ST segment was downsloping; if the number of major vessels colored by fluoroscopy was small. It was also indifferent between the presence of heart attack with the fasting blood sugar >120mg/dl as well as with the resting blood pressure. Besides that, we were able to observe that subjects that were female, subjects that had ST-T wave abnormality T wave inversions and/or ST elevation or depression of > 0.05 mV as well as subjects that had non-typical angina were more prone to higher chance of getting a heart attack.

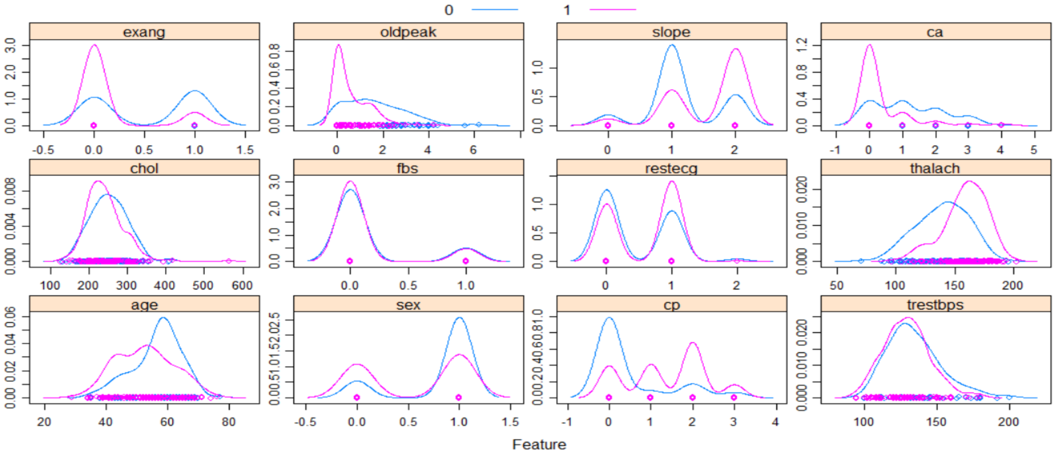
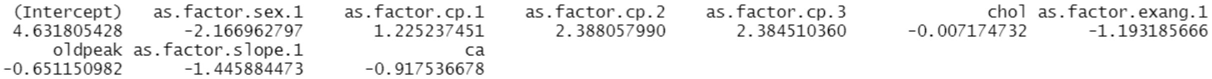


Figure 2: *Density plots of the response variable, presence of heart attack versus the 12 predictors, exang, oldpeak, slope, ca, chol, fbs, restecg, thalach, age, sex, cp, and trestbps.*

Additional plots and discoveries were included in Appendix A.

1.3 - STATISTICAL ANALYSIS / MODELING

We started off using the training dataset to fit the model with Conventional Logistic by including all of the 12 predictors with its dummy variables (Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, and Ca) and we ran the model selection process which in this case, we chose stepwise regression since it did both backward elimination and forward selection. In the end, the model consisted of 9 predictors with an Akaike Information Criterion (AIC) value of 170.33 as shown in Figure 5.

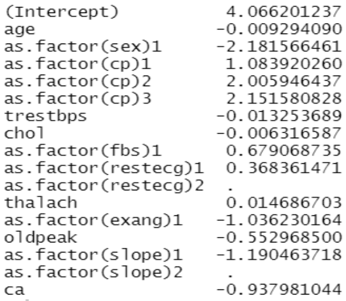


```
  (Intercept)   as.factor.sex.1    as.factor.cp.1    as.factor.cp.2    as.factor.cp.3              chol as.factor.exang.1
  4.631805428      -2.166962797       1.225237451       2.388057990       2.384510360      -0.007174732      -1.193185666
      oldpeak as.factor.slope.1                ca
 -0.651150982      -1.445884473      -0.917536678
```

Figure 5: *This figure shows the coefficients of the 9 covariates (sex, cp1, cp2, cp3, chol, exang1, oldpeak, slope1 and ca) in the model along using the conventional logistic regression.*

We also further assessed the Hosmer-Lemeshow goodness of fit of the model. Since the p-value was 0.8911, thus, it was not significant at alpha level 0.05 and we were able to conclude that there was no lack of fit issue in the model. Beyond that, we evaluated the model diagnostic and based on the residuals versus fitted values plot, we were able to observe that most of the residuals were located close to the loess line and the overall fit of residuals was good as no obvious pattern was found. Furthermore, the model prediction accuracy using the test dataset resulted in a correct classification rate of 79.77528% which 0.5 was used as the classification threshold and the area under the curve (AUC) of the ROC curve was relatively high, approximately 0.9005102.

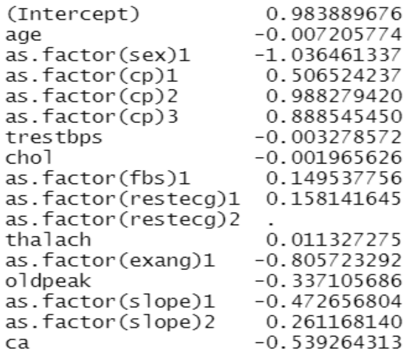
Aside from fitting the training dataset using the conventional logistic regression method, we also fitted the model using the logistic regression with Lasso penalty. Using the L1 shrinkage, we got the best lambda that gave the highest AUC to be 0.00171024. As shown in figure 6, the coefficient for the variables resting electrocardiographic that showed probably or definite left ventricular hypertrophy by Estes' criteria (restecg = 2) and downsloping slope of the peak exercise ST segment (slope = 2) was shrunk to 0. The test set prediction accuracy using the fitted model with Lasso resulted in a correct classification rate of 82.02247% which was higher than the conventional logistic regression. The AUC of the ROC curve was roughly 0.9096939.



```
(Intercept)           4.066201237
age                  -0.009294090
as.factor(sex)1      -2.181566461
as.factor(cp)1        1.083920260
as.factor(cp)2        2.005946437
as.factor(cp)3        2.151580828
trestbps             -0.013253689
chol                 -0.006316587
as.factor(fbs)1       0.679068735
as.factor(restecg)1   0.368361471
as.factor(restecg)2   .
thalach               0.014686703
as.factor(exang)1    -1.036230164
oldpeak              -0.552968500
as.factor(slope)1    -1.190463718
as.factor(slope)2     .
ca                   -0.937981044
```

Figure 6: *This figure shows the coefficients of the 14 covariates where some levels of the predictors were shrunk to 0 after fitted using the logistic regression with Lasso penalty.*

Apart from that, we had also fitted the training dataset with Elastic Net where it combined the L1 and L2 penalties of the Lasso and Ridge methods. In order to find the best alpha that gave us the highest prediction accuracy or in other words, the lowest mean square error(MSE), we first chose a sequence of alphas and for each alpha we fitted the elastic net regression into the training set. The alpha that corresponded to the lowest MSE turned out to be 0.2222222. We then proceeded to fit the elastic net regression with alpha=0.2222222 and the best lambda turned out to be 0.05958795 with a test set prediction result of 84.26966% while the test set performance based of the AUC of the ROC curve was about 0.9081633. As shown in figure 7, only one predictor's coefficient (restecg =2) was shrunk to 0 and 12 predictors were still kept in the model.



```
(Intercept)           0.983889676
age                  -0.007205774
as.factor(sex)1      -1.036461337
as.factor(cp)1        0.506524237
as.factor(cp)2        0.988279420
as.factor(cp)3        0.888545450
trestbps             -0.003278572
chol                 -0.001965626
as.factor(fbs)1       0.149537756
as.factor(restecg)1   0.158141645
as.factor(restecg)2   .
thalach               0.011327275
as.factor(exang)1    -0.805723292
oldpeak              -0.337105686
as.factor(slope)1    -0.472656804
as.factor(slope)2     0.261168140
ca                   -0.539264313
```

Figure 7: *This figure shows the coefficients of the 15 covariates where only the predictor (restecg = 2) was shrunk to 0 after fitted using the elastic net regression.*

1.4 CONCLUSION / DISCUSSION

In conclusion, for the presence of heart disease dataset, our best final model chosen was the model fitted with conventional logistic regression selected by the stepwise selection since it had the least number of predictors and the accuracy of the test dataset and and area under the curve (AUC) of the ROC curve were comparable to the LASSO and Elastic Net Regression.

	Conventional Logistic	**Lasso**	**Elastic Net**
Test set correct classification rate	**0.7977528**	0.8202247	0.8426966
AUC for Test data	**0.9005102**	0.9096939	0.9081633
Number of Variables	**9**	14	15

Table 4: *Comparison of the test set correct classification, area under the curve and number of variables in the model*

The formula of the best final model and its interpretation was shown as per below:
$\text{Logit}(\pi) = 4.6318 - 2.167(\text{sex}=1) + 1.2252(\text{cp}=1) + 2.388(\text{cp}=2) + 2.3845(\text{cp}=3) - 0.00717(\text{chol}) - 1.1932(\text{exang}=1) - 0.651(\text{oldpeak}) - 1.4459(\text{slope}=1) - 0.9175(\text{ca})$

The odds of having a heart attack decreased by a factor of 0.1145 for male subjects compared to the odds for female subjects given all other predictors unchanged. Apart from that, given that all other predictors were unchanged, the odds of having a heart attack increased by a factor of 3.405 for subjects who had atypical angina chest pain, increased by a factor of 10.892 for subjects who had non-anginal pain, and increased by a factor of 10.854 for subjects who had asymptomatic pain as compared to the odds for subjects who had typical angina chest pain. The odds of having more chances of heart attack decreased by a factor of 0.9929 for every increase in mg/dl in serum cholesterol. Besides that, the odds of having a heart attack decreased by a factor of 0.3033 for subjects who had exercise induced angina compared to the odds for subjects who had no exercise induced angina while the odds of having a heart attack decreased by a factor of 0.5214 for every increase in subjects' ST depression induced by exercise relative to rest. Finally, the odds of having a heart attack decreased by a factor of 0.236 for the flat slope of the peak exercise ST segment compared to the odds for upsloping slope of the peak exercise ST segment while the odds decreased by a factor of 0.3995 for every increase in number of major vessels (CA) colored by fluoroscopy given all other predictors unchanged.