The Art of Heart Beats

—— Cleveland Database Research

In this article, we aim to analysis the Cleveland's heart disease dataset to interpret the relationships between 14 listed variables, then use analytical models to exam the accuracy of each predictions.

Data Exploration

1. Missing value treatment

At first, there are 6 missing values, which is relatively small to the sample size 303, so we delete them. The total sample size decreases from 303 to 297.

2. Variable Identification

All data types of variables are numeric. The category of variables is: categorical (2. Sex, 3. Cp, 6. Fbs, 7. Restecg, 9. Exang, 11. Slope, 12. Ca, 13. Thal, 14. Num) and continuous (1. Age, 4. Trestbps, 5. Chol, 8. Thalach, 10. Oldpeak).

2.1. Univariate Analysis

For the variable age, the elements are mostly in range 40-70 years old; meanwhile, the number of female data and male data are 96 and 201.

2.2. Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at 0.05 significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Continuous & Continuous

We use scatter plot to do bi-variate analysis between two continuous variables. Among continuous variables, we find that there is clear negative linear relationship between age and thalach (Shown as Exhibit 1). Trestbps and thalach are relatively independent. Chol and thalach are relatively indenpent Exhibit 1, etc(Shown as the black √in Exhibit 2).

Categorical & Categorical

We use Chi-Square Test to find the relationship between two categorical variables. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. When p-value is quite small (smaller than 0.05), we reject the hypothesis that two variables are independent. That is to say, when p-value is smaller than 0.05, the two variables are correlated. According to our R results, we conclude correlated pairs (Shown as the green √in Exhibit 2)

Categorical & Continuous

Comparing to categorical and continuous variables, we use ANOVA to test that relationship of categorical and continuous variables. Similarly, when p-value is smaller than 0.05, the two variables are correlated. Then we got correlated pairs. (Shown as the red √in Exhibit 2)

3. Outlier Treatment

According to our observation, we find some observations that appear far away and diverges from an overall pattern which can result in wildly wrong estimations, so we conclude those must be outliers, so remove the outliers of variables age, trestbps, chol, thalach, oldpeak.(An example of trestbps is shown in exhibit 3)

Hypotheses- Methodologies- Results and Conclusion

1. Use other 13 variables to predict num

1.1. Logistic Regression

Hypotheses: As we can see the correlation between num and other variables, we assume that variables like age, sex, cp, trestbps, restecg, thalach, exang, oldpeak, ca and thal have significant effect on num while other variables are nonsignificant.

Methodologies: We use logistic regression to test our hypothesis. At first, we turn the value of num 1,2,3, or 4 into 1, then we define "num" as response and all other 13 variables as predictors. We use AIC to decide the best model and remove nonsignificant variables (age, restecg, thalach, fbc and exang) from the original model. The formula is shown in exhibit 5. And then, we split the data into training and test data and get 0.8345324 accuracy.

Results and Conclusion: The result of logistic regression is different from our correlation analysis (shown as exhibit 4). Maybe it is because the removed data like age, restecg, thalach, fbc and exang have correlation with factors in the relative best model which make them nonsignificant, like exercise induced angina may cause chest pain. As to interpret the best model, we found some supporting facts, for example, ST depression induced by exercise relative to rest has effect on diagnosis of heart disease.

1.2. Decision tree

Then we use decision tree to predict num before removing the factors, then we get the decision tree (shown as Exhibit 6) with 0.8057554 accuracy. We also use cross-validation and pruning to obtain smaller trees (shown as Exhibit 7), with the same prediction accuracy.

2. Thalach, exang, oldpeak, slope, ca, thal and num to predict CP

Hypotheses: Chest pain is common in the US. Different types of chest pain have distinct causes, and some might be serious or even life-threatening. We hope to find factors that affect or decide the types of chest pain so that people can distinguish different types better and deal with it properly. Based on the bivariate analysis above (shown as exhibit 2), we assume thalach, exang, oldpeak, slope, ca, thal and num influence cp.

Methodologies: [Multiple Linear Regression]We run a multiple linear regression with cp as the output variable and thalach, exang, oldpeak, slope, ca, thal and num as the input variables. Since there are only seven predictors, we use the best subset selection method, rather than forward or backward selection method, to pick the best model with the highest R^2. According to the Exhibit 12, the model is:

$$\text{cp} = 3.518 - 0.006 \times \text{thalach} + 0.440 \times \text{exang} + 0.036 \times \text{ca} + 0.033 \times \text{thal} + 0.162 \times \text{num}$$

The p-value for the model is practically 0, which shows the model is significant. However, the adjusted R^2 is 0.2118, which means only 21.18% of the variability in cp can be explained using the independent variables of the model. The model doesn't work well.

Using inappropriate regression might be the reason for the low fitness of the model. Multiple linear regression assumes the numerical amount in dependent variable should be the meaning data point. But the dependent variable cp is a categorical variable. To get a more fitted model, we choose the KNN algorithm, whose dependent variable can be categorical.

[KNN algorithm] First, we split data into two subsets, one training data and one test data. Then, we run the KNN with k=1, 2, 3, 4, 5, 6, 7, 8, 9. Finally, we pick the best model based on the accuracy of predictions for the test data.

Results and Conclusion: When k=8, the model is relatively better. However, the model still doesn't work well and only 51.8% of the prediction is accurate. We think there might be three reasons.

- **The regression method is still not suitable.** Multinomial logistic regression might be useful.
- **Some significant predictors are not included in the model.** On the one hand, we ignore some internal variables. Our regression is based on the result of the bivariate analysis. However, this

analysis does not consider the suppression effect of other variables. That is, sometimes the confounding effect of C on A and B might incidentally cancel out the effect of B on A. For example, American Heart Association argues the signs of chest pain in women may be different from those in men.[1] The men are more likely to suffer typical angina. However, affected by other variables, this relationship is not shown in the bivariate analysis. On the other hand, our data mainly focus on heart-related factors, and some external variables are not taken into account. For example, the difference in chest wall tenderness will influence chest pain type.[2]

- **Irrelevant variables are taken into the model due to attribution error.** Regression tests the causal relationship between variables while bivariate analysis shows the correlation. The variables sharing the same correlation or tendencies cannot prove they have a causal relationship. Sometimes, it is just a coincidence.

3. Unsupervised Learning:

Except the supervised learning, we also want to analysis the dataset with unsupervised learning. Before studying the input variables, we found that the elements of input variables are internally imbalance. So, we use scale to standardize the data before starting the cluster analysis.

Hypotheses& Methodologies: First, we don't clearly know the criteria for clusters partitioning. So we first decided to use the Hierarchical Clusters method to process the overall data to find interesting phenomena.

[Hierarchical Clusters: 1st Round] We took all the variables into account and divide the data into two clusters. Plot the clusters by two features then get 78 plots. 4 of them have significant distribution characteristics (as shown as Exhibit 9). Then we find sex and age are important factor to decide which cluster the data belong to.

[Hierarchical Clusters: 2ed Round]A 1982 research article *"Physical attractiveness and blood pressure: Sex and age differences" (Hansell 1982)*, which was published six years earlier than the data set, pointed out that the relatively unattractive young woman's blood pressure is higher than the attractive young female or general male's blood pressure, while the older men's and women's blood pressure tends to be equal. Combining the article with our four findings in round 1 data processing, we cross compare the variables age, sex and resting blood pressure and plot the graph to shows significant relationship between the variables.

Results and Conclusion: For round 2 analysis, we found that although both age and sex generate two clusters for resting blood pressure, both age and sex will become unimportant when compared with resting blood pressure. (Exhibit 10, 11) This result does not match the results of the 1982 article. Through our analysis, we conclude that there are mainly three explanations for the inconsistency.

- Psychological changes occurred after six years, and the difference in blood pressure gradually disappeared.
- The female data in our data is initially less than that of men, so it creates inconsistency.
- There is a vast difference between the female data in the Cleveland region and the data for women across the country.

Ending

In general, we analyzed the data in different ways, some of which were significant and some failed. In general, the Cleveland database sample size is still relatively small. We need to analyze with larger samples in future research.

[1] www.heart.org. (2019). Angina in Women Can Be Different Than Men. American Heart Association.
[2] Constant, J. (1983). The clinical diagnosis of nonanginal chest pain: The differentiation of angina from nonanginal chest pain by history. Clinical Cardiology, 6(1), pp.11-16.

Appendix

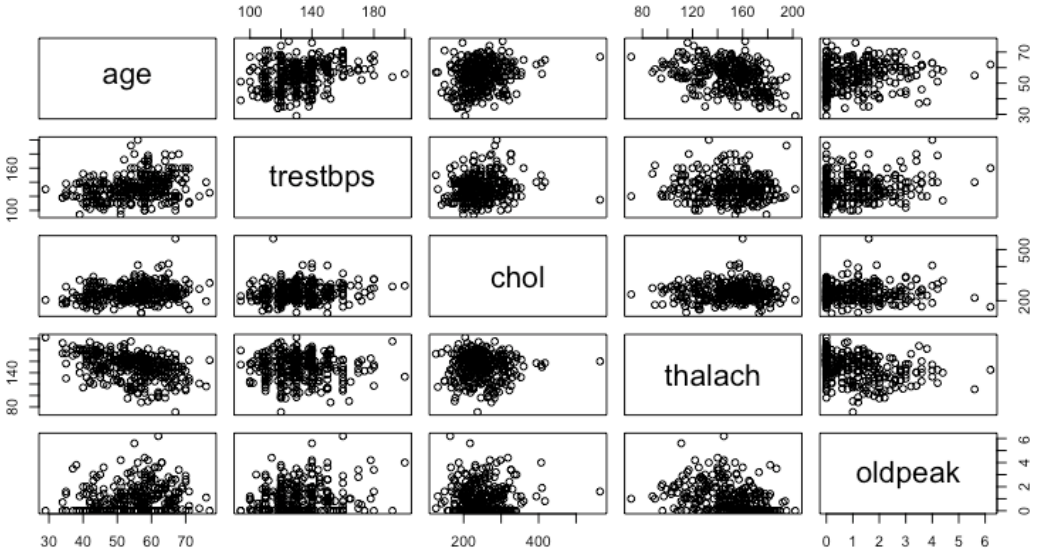

Exhibit 1

	1.age	2.sex	3.cp	4.trestbps	5.chol	6.fbs	7.restecg	8.thalach	9.exang	10.oldpeak	11.slope	12.ca	13.thal	14.num
1.age						✓	✓	✓			✓	✓	✓	✓
2.sex					✓				✓				✓	✓
3.cp								✓	✓	✓	✓	✓	✓	✓
4.trestbps						✓	✓				✓		✓	✓
5.chol		✓					✓					✓		
6.fbs	✓			✓								✓		
7.restecg	✓			✓	✓						✓			✓
8.thalach	✓		✓						✓		✓	✓	✓	✓
9.exang		✓	✓					✓		✓	✓	✓	✓	✓
10.oldpeak			✓						✓		✓	✓	✓	✓
11.slope	✓		✓	✓			✓	✓	✓	✓			✓	
12.ca	✓		✓		✓	✓		✓	✓	✓			✓	✓
13.thal	✓	✓	✓	✓				✓	✓	✓	✓	✓		✓
14.num	✓	✓	✓	✓			✓	✓	✓	✓		✓	✓	

Exhibit 2

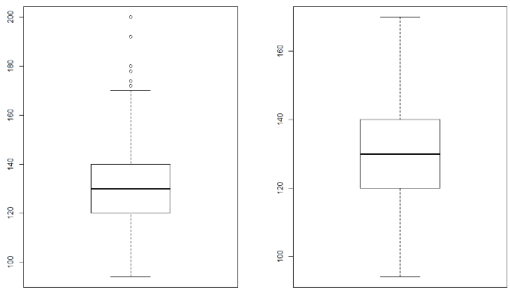

Exhibit 3

	14.num	LR
1.age	✓	
2.sex	✓	✓
3.cp	✓	✓
4.trestbps	✓	✓
5.chol		✓
6.fbs		
7.restecg	✓	
8.thalach	✓	
9.exang	✓	
10.oldpeak	✓	✓
11.slope		✓
12.ca	✓	✓
13.thal	✓	✓
14.num		

Exhibit 4

$$p(sex1, cp1, cp2, cp3, cp4, trestbps, chol, oldpeak, slope2, slope3, ca1, ca2, ca3, thal6, thal7)$$

$$= \frac{e^{-10.29+1.35*sex1+1.10*cp2+0.023*cp3+2.37*cp4+0.022*trestbps+0.0084*chol+0.60*oldpeak+1.25*slope2+0.85*slope3+2.30*ca1+2.62*ca2+2.13*ca3+0.18*thal6+1.69*thal7}}{1+e^{-10.29+1.35*sex1+1.10*cp2+0.023*cp3+2.37*cp4+0.022*trestbps+0.0084*chol+0.60*oldpeak+1.25*slope2+0.85*slope3+2.30*ca1+2.62*ca2+2.13*ca3+0.18*thal6+1.69*thal7}}$$

Exhibit 5

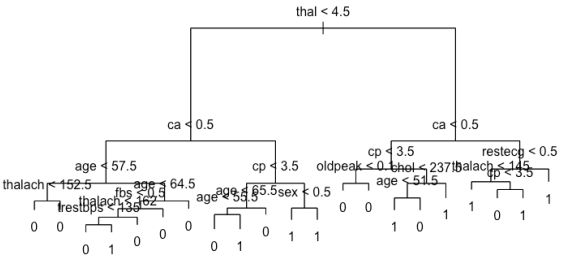

Exhibit 6

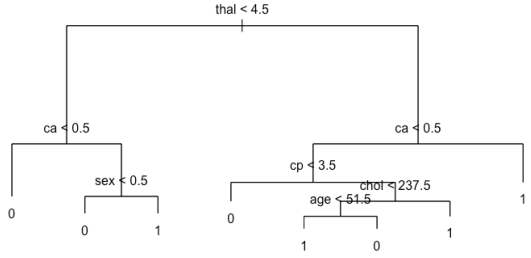

Exhibit 7

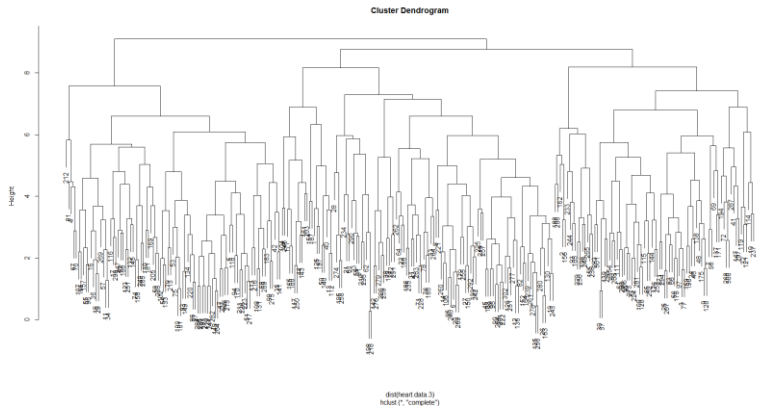

Exhibit 8

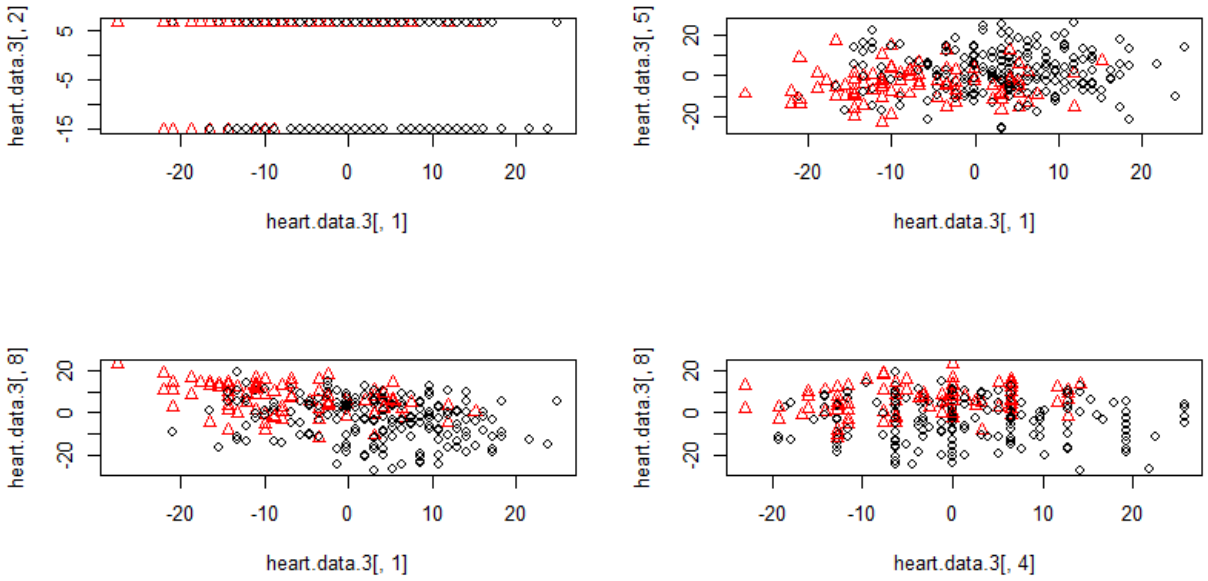

Exhibit 9

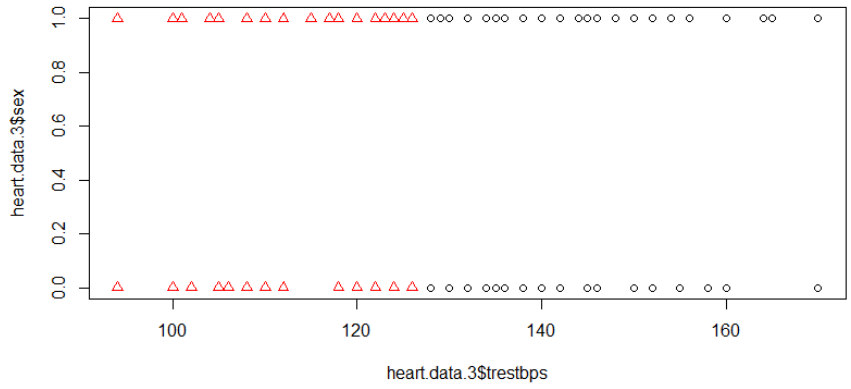

Exhibit 10

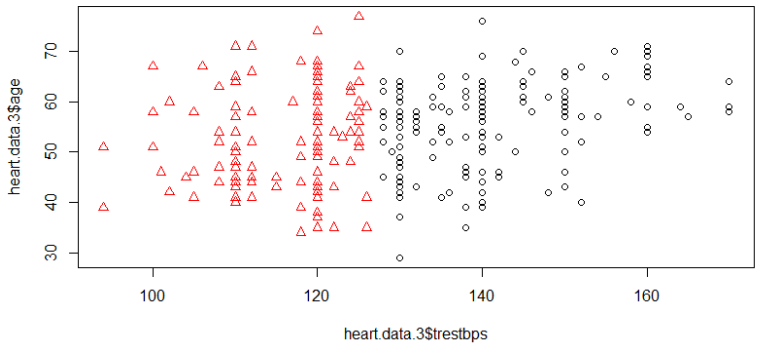

Exhibit 11

```
Call:
lm(formula = cp ~ thalach + exang + ca + thal + num, data = heart.data.4)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8101 -0.5759  0.2099  0.4635  1.4303

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.517547   0.456446   7.706 2.43e-13 ***
thalach     -0.005626   0.002644  -2.128 0.034262 *
exang        0.439849   0.127325   3.455 0.000639 ***
ca           0.035764   0.066058   0.541 0.588674
thal         0.032858   0.031797   1.033 0.302355
num          0.161661   0.059178   2.732 0.006712 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8603 on 272 degrees of freedom
Multiple R-squared:  0.226,     Adjusted R-squared:  0.2118
F-statistic: 15.89 on 5 and 272 DF,  p-value: 9.934e-14
```

Exhibit 12