Car Price Prediction
and
Heart Disease Classification

Hamid Hamidi, Thet Nyein, and Yanzhao Qian

Authors are in alphabetic order and have equal contribution.

Contents

Contents

Abstract

Car Price Prediction

Introduction

Data Collection and exploration

Statistical Analysis

Conclusion

Heart Disease Classification

Introduction

Data Collection and exploration

Statistical Analysis

3 GLM models and stepwise AIC selection

To begin with, let's assign each feature with a variable name.

Data Dictionary			
Variable Name	Definition	Explanation	Variable name
Age	age of the patient	years	x_1
Sex	sex of the patient	M: Male, F: Female	x_2
ChestPainType	chest pain type	TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic	x_3
RestingBP	resting blood pressure	mm Hg	x_4
Cholesterol	serum cholesterol	1: if FastingBS $>$ 120 mg/dl, 0: otherwise	x_5
FastingBS	fasting blood sugar	Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria	x_6
RestingECG	resting electrocardiogram results	Numeric value between 60 and 202	x_7
MaxHR	maximum heart rate achieved	Y: Yes, N: No	x_8
ExerciseAngina	exercise-induced angina	Numeric value measured in depression	x_9
Oldpeak	oldpeak = ST	Numeric value measured in depression	x_{10}
ST_Slope	the slope of the peak exercise ST segment	Up: upsloping, Flat: flat, Down: downsloping	x_{11}
HeartDisease	output class	1: heart disease, 0: Normal	y

Modeling All the Main Covariates

Since the response is having heart disease or not, it is a binomial distributed response. So it is suggested we could use a logistic regression model. At first, we created a logistic model of all the main effects, and the

formula of the model is

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i$$

where p is the probability to get the heart disease, and $\beta_i, i = 1, 2, \ldots, 11$ is the coefficients of the parameter, β_0 is the intercept. So we fitted the model, and got the estimated parameters shown in the table below.

	Estimate	Pr(>|z|)	significance
(Intercept)	-1.16	0.411	not significant
Age	0.0166	0.21	not significant
SexM	1.47	1.6e-07	***
ChestPainTypeATA	-1.83	2.03e-08	***
ChestPainTypeNAP	-1.69	2.34e-10	***
ChestPainTypeTA	-1.49	0.00058	*
RestingBP	0.00419	0.485	not significant
Cholesterol	-0.00411	0.000154	*
FastingBS	1.14	3.59e-05	**
RestingECGNormal	-0.177	0.515	not significant
RestingECGST	-0.269	0.443	not significant
MaxHR	-0.00429	0.393	not significant
ExerciseAnginaY	0.9	0.000231	*
Oldpeak	0.381	0.00131	.
ST_SlopeFlat	1.45	0.000703	*
ST_SlopeUp	-0.994	0.0272	.

As we can see from the table, there are some variables that are not significant. We should drop some variables to make the model simpler. We choose backward step selection to do so. And we get our estimated paraeters shown in the table below.

	Estimate	Pr(>|z|)	significance
(Intercept)	-1.72	0.0436	.
Age	0.0231	0.0518	not significant
SexM	1.47	1.36e-07	***
ChestPainTypeATA	-1.86	8.89e-09	***
ChestPainTypeNAP	-1.72	6.13e-11	***
ChestPainTypeTA	-1.49	0.000494	*
Cholesterol	-0.00398	0.000106	*
FastingBS	1.13	3.41e-05	**
ExerciseAnginaY	0.936	8.21e-05	**
Oldpeak	0.377	0.00121	.
ST_SlopeFlat	1.46	0.000654	*
ST_SlopeUp	-1.03	0.0211	.

From the table, we can see all the variables are significant now. here is the **model 1** formula.

$$\begin{aligned}\log \frac{\hat{p}}{1-\hat{p}} = &-1.7 + 0.023\text{Age} + 1.5\text{SexM} - 1.9\text{ChestPainTypeATA} \\ &- 1.7\text{ChestPainTypeNAP} - 1.5\text{ChestPainTypeTA} - 0.0040\text{Cholesterol} \\ &- 1.1\text{FastingBS} \ + 0.94\text{ExerciseAnginaY} \ + 0.38\text{Oldpeak} \\ &+ 1.5\text{ST\_slopeFlat} - 1.0\text{ST\_slopeUp}\end{aligned}$$

We need to test if the selected model is good enough to represent the origin mode. We did log-likelihood ratio test for the step-wise selected variables to see if the drop out is good. The null hypothesis is

$$H_0 : \beta_{\text{RestingBP}} = \beta_{\text{RestingECG}} = \beta_{\text{MaxHR}} = 0 \text{ v.s } H_1 : \text{At least one of these parameters not 0}$$

We using the formula to get LLR statistic as

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-297.0925 + 297.9042) = 0.804$$

with degrees of freedom of 4. So we can calculate the p-value is 0.8046016, which is very high. So we cannot reject H_0. So we can accept the reduced model.

Modeling the Square of Numerical Variables Now we investigated the square of the numerical variables. We did this approach because response may have some quadratic effect of the numerical variables, and square of categorical variables do not make any difference. The odds model is

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i + \beta_{12} x_1^2 + \beta_{42} x_4^2 + \beta_{72} x_7^2 + \beta_{102} x_{10}^2$$

The model is fitted by R. And all the parameters and significance of them are demonstrated in the table below.

	Estimate	Pr(>|z|)	significance
(Intercept)	3.34e+00	4.96e-01	not significant
Age	-2.90e-02	7.86e-01	not significant
SexM	1.49e+00	2.00e-07	***
ChestPainTypeATA	-1.71e+00	2.00e-07	***
ChestPainTypeNAP	-1.68e+00	0.00e+00	***
ChestPainTypeTA	-1.38e+00	2.31e-03	.
RestingBP	-2.86e-02	5.47e-01	not significant
Cholesterol	-1.16e-02	2.87e-05	**
FastingBS	1.08e+00	1.73e-04	*
RestingECGNormal	-1.65e-01	5.53e-01	not significant
RestingECGST	-2.96e-01	4.10e-01	not significant
MaxHR	-2.29e-02	5.81e-01	not significant
ExerciseAnginaY	1.08e+00	2.09e-05	**
Oldpeak	-3.04e-01	3.33e-01	not significant
ST_SlopeFlat	1.77e+00	1.12e-04	*
ST_SlopeUp	-7.29e-01	1.26e-01	not significant
I(MaxHR^2)	7.45e-05	6.24e-01	not significant
I(Age^2)	4.72e-04	6.37e-01	not significant
I(Oldpeak^2)	2.67e-01	1.96e-02	.
I(RestingBP^2)	1.20e-04	4.95e-01	not significant
I(Cholesterol^2)	2.17e-05	2.22e-03	.

But there are still some useless covariates in the model. To simplify the model, a backward step-wise selection was used to the square of numerical model. The parameters and their significance of selected model are shown in the table.

	Estimate	Pr(>|z|)	significance
(Intercept)	-1.05e+00	1.25e-01	not significant
SexM	1.50e+00	1.00e-07	***
ChestPainTypeATA	-1.72e+00	1.00e-07	***
ChestPainTypeNAP	-1.68e+00	0.00e+00	***
ChestPainTypeTA	-1.37e+00	1.98e-03	.
Cholesterol	-1.18e-02	9.90e-06	***
FastingBS	1.05e+00	2.10e-04	*
ExerciseAnginaY	1.02e+00	1.92e-05	**
ST_SlopeFlat	1.74e+00	1.20e-04	*
ST_SlopeUp	-7.23e-01	1.22e-01	not significant
I(Age^2)	2.22e-04	4.77e-02	.
I(Oldpeak^2)	1.70e-01	2.21e-04	*
I(Cholesterol^2)	2.24e-05	1.44e-03	.

We now apply Log-likelihood ratio test to get

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-286.9175 + 288.5404) = 3.245983$$

So the p-value is 0.9179859, which means the reduction is highly possible.

Modeling the Interaction of Numerical We also modeled the interaction between numerical terms, and the model is

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i + \sum_{i \in \mathcal{N}} \beta_{i2} x_i^2 + \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} \beta_{ij} x_i x_j$$

We also did backward stepwise selection to reduce the model, ane got all the coefficients demonstrated in the table below

	Estimate	Pr(>|z|)	significance
(Intercept)	-1.950000	1.29e-01	not significant
Age	0.021300	9.17e-02	not significant
SexM	1.420000	4.00e-07	***
ChestPainTypeATA	-1.810000	0.00e+00	***
ChestPainTypeNAP	-1.660000	0.00e+00	***
ChestPainTypeTA	-1.450000	9.97e-04	*
Cholesterol	-0.003860	2.71e-04	*
FastingBS	1.200000	1.44e-05	**
MaxHR	0.000751	8.97e-01	not significant
ExerciseAnginaY	0.993000	6.60e-05	**
Oldpeak	0.668000	3.73e-01	not significant
ST_SlopeFlat	1.810000	7.59e-05	**
ST_SlopeUp	-0.732000	1.20e-01	not significant
I(Oldpeak^2)	0.324000	6.27e-03	.
MaxHR:Oldpeak	-0.007750	1.42e-01	not significant

Analysis of the 3 Models

1. ROC, Deviance Test, Pseudo-R square,

Conclusion

References

Appendix