Car Price Prediction and Heart Disease Classification

Hamid Hamidi, Thet Nyein, and Yanzhao Qian

Authors are in alphabetic order and have equal contribution.

Contents

Abstract 3

Car Price Prediction 4

Introduction 4
Data Collection and exploration 4
Statistical Analyses 5
Conclusion 5

Heart Disease Classification 6

Introduction 6
Data Collection and exploration 7
Statistical Analyses 7
Three GLM models and stepwise AIC selection 7
Analysis of the Three Models 11
Conclusion 11

Appendix 11

Appendix Figures 12

References 13

Abstract

Generalized Linear Models (GLMs) are the extension of the ordinary linear regression models. GLMs enable us to use different distributions for the response with distinctive link functions [1]. Here, we use two different data sets to show the broad applications of the GLMs in real-world problems, one of which is the “Car Price Prediction” [2], and the other is “Heart Failure prediction” [3]. In our analyses, we focus on model fitting and highlighting the most important variables instead of predicting desired outcomes and their accuracy. Our study of each data set is reported in its corresponding section. In the following, we discuss why we have chosen these data sets and provide a detailed description of our analyses along with the reasons and intuitions behind them. In both of these studies, all analyses were performed using the R programming language [4].

Car Price Prediction

Introduction

One of the largest automotive markets in the world is the USA car market [5]. Since 1982, when Honda invested in the USA car market, many other companies have been joining and competing in the USA car market resulting in foreign investment of more than \$110 billion [5]. These days, with skilled workers, local and governmental supports, a huge consumer market, and many other reasons, the USA car market is a primer market in the car industry. A new Chinese car company wants to join and compete in the USA car market. In the following, our goal is to identify significant variables affecting the car price and quantify their significance. These analyses are usually performed by a third party, such as a consulting company, or the business strategy division of the investing company. According to our findings, they can manipulate many variables, such as the car design, to have a better business strategy to enter the USA car market. These analyses can directly affect the success of billions of dollars investment. Consequently, our analyses are vital and should be detailed and valid.

We found out the car price (response) distribution is quite close to the Gamma distribution; therefore, we used the GLM with Gamma distribution and logarithmic link function to model the price of cars for distinctive variables. We also suspected that it might be possible to model the logarithm of price with Gaussian distribution and identity link function. However, the distribution of the logarithmic price is not close to the Normal distribution. Consequently, we only used the Gamma distribution with the logarithmic link function. We performed variable selection and selected the most reasonable model (details in the Statistical Analyses section).

Using these analyses, we were able to identify several significant variables contributing to the car price, such as the car manufacturer (or the so-called brand of the car), the engine location (cars with rear engines are usually sports cars with higher prices), and the engine size (the bigger the higher the price).

Our data set, and consequently, our analyses have some limitations as well. For instance, electric cars are more than 2.5% of the USA car market [6] but are not included in our data set. Additionally, luxury brands such as Rolls-Royce and Lincoln are missing. Furthermore, the majority of sports cars are missing in our data set, showing our limitation in analyzing the sports and luxury car price variables. In the following, we present a detailed description of our analysis, methods, and results.

Data Collection and exploration

The data were collected from the Kaggle website (kaggle.com), an online open-source community of data scientists and machine learning practitioners. One can easily access the online version of our data through [2]. The data did not contain any missing values and was ready for analysis. However, we made minor changes and corrections in the data set.

We removed the CAR ID column as it does not contain useful information for our analyses. Additionally, we changed the names of the cars into manufacturers' names. This way, the variable would represent the car brand (or manufacturer) reputation, which might have an impact on car price, instead of the model of the car, which is unique for most cars and would not impact the car price.

Afterward, we tried to figure out the response distribution to use the appropriate link function and family of distribution [1]. The distribution of the response resembles the Gamma distribution (Figure 1 (A)). We also visualized the logarithm of the response since it might be Gaussian (Figure 1 (B)). As one can see in Figure 1 (B), the logarithm of price does not resemble the Gaussian distribution. Therefore, we decided to only use the Gamma distribution with the log link function (See Statistical Analyses for more details).

Figure 2 shows an overview of the manufacturers, the range of their cars' prices, and the fuel type of their productions. As it is obvious, electric cars are missing, and diesel cars are the minority. Moreover, as shown in Figure 2, the car brands (or manufacturers) may affect the car price. For instance, cars from Porsche have a higher price compared to cars from Nissan or Mazda. Furthermore, we can see that famous sports car brands such as Ferrari and luxury manufacturers, such as Rolls-Royce and Lincoln, are missing.

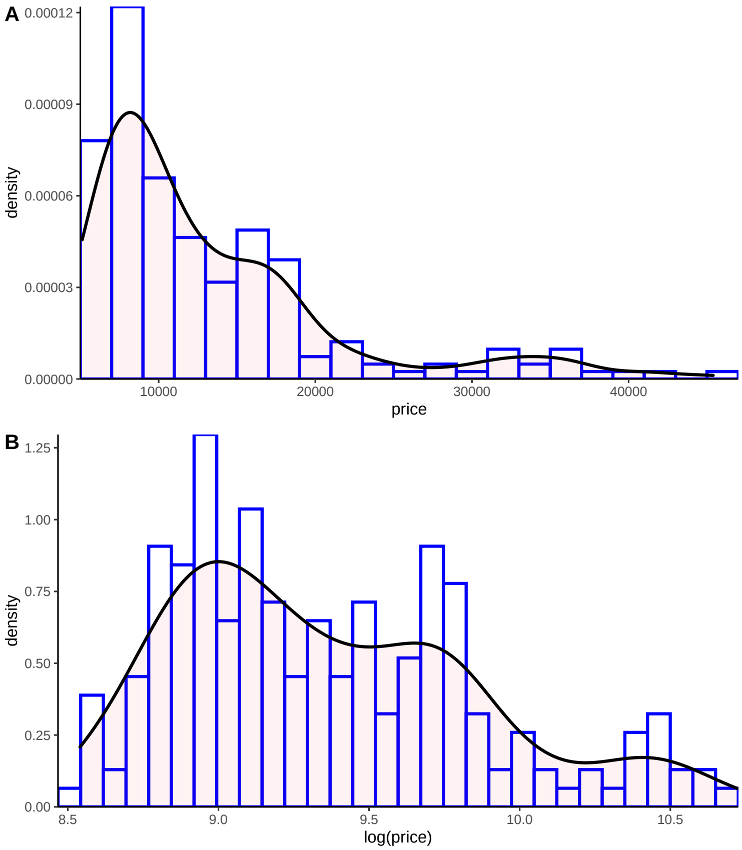


Figure 1: Distribution of the price.
(A) The distribution of the response (B) The distribution of the logarithm of the response.

In Appendix Figure 1, we can see that the car price is correlated with engine size; however, it might not be a linear correlation. Also, what stands out in this figure is the general growth of the engine size with increasing the number of Cylinders, and also, the response will rise with increasing any of them.

We also suspected that there might be some trends with the quadratic increase of numerical variables. Therefore, we investigated these patterns. For instance, in Appendix Figure 2, we divided the wheelbase[1] of cars into four groups and visualized the trend between the wheelbase and the response. As this figure shows, the car price is not growing linearly with the increase of the wheelbase. Hence, we included quadratic terms in our statistical model as well. In the next section, the details of our statistical analyses are described.

Statistical Analyses

Conclusion

[1] In cars, the wheelbase is the distance between the front and rear wheels [7].

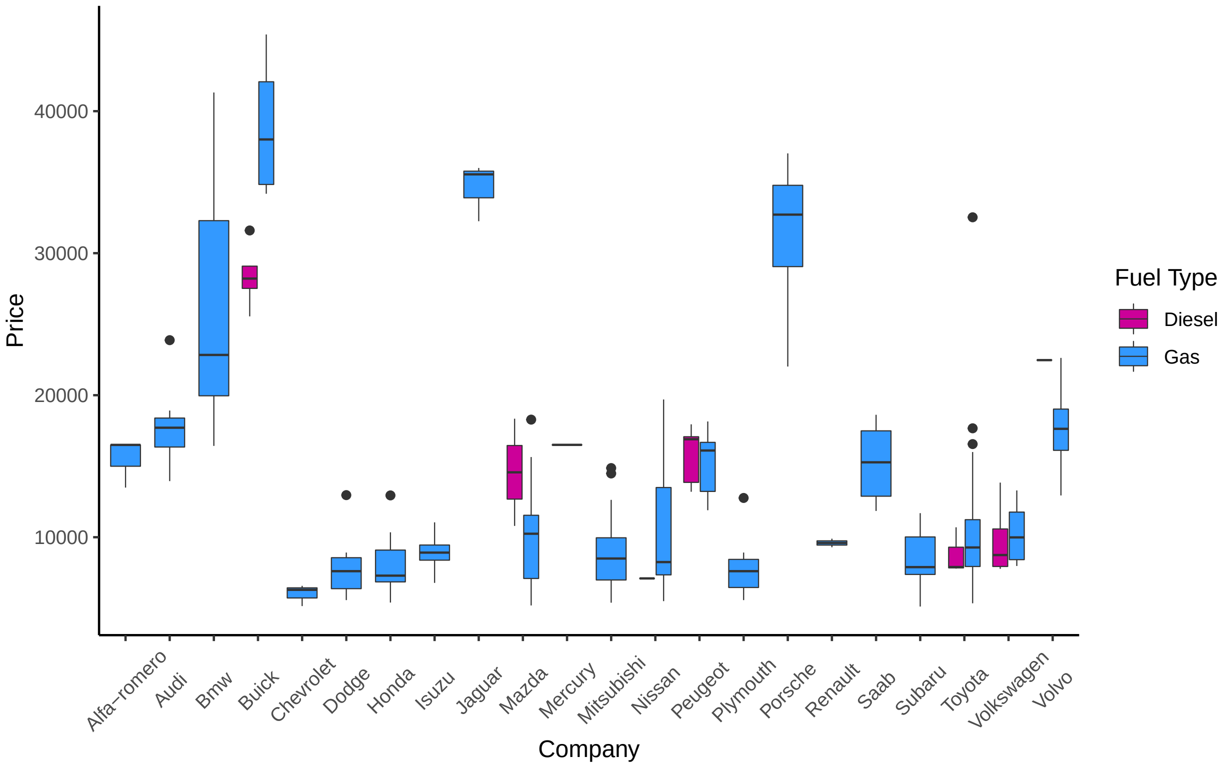


Figure 2: The range of car price in different brands and fuel types

Heart Disease Classification

Introduction

Heart Failure, also known as congestive heart failure, can broadly be defined as a condition that happens when the heart cannot supply the body's need for Oxygen and blood [8]. According to the latest annual statistical report from the American Heart Association and the National Institutes of Health, about 6.2 million adults in the United States have heart failure [9]. Furthermore, in 2018, heart failure was mentioned on 379,800 death certificates (13.4%) [9] and cost about $30 billion annually [10].

This suggests that identifying the core health behaviors and risk factors influencing heart failure is critical not only for our community health but also for our economy. Therefore, we decided to analyze the "Heart Failure prediction" data set [3] to find variables playing a key role in heart failure. As the response in our data set is binary (0 or 1), we used logistic regression to model the probability of having heart failure. Moreover, we performed variable selection to select the best model and determine major factors in heart failure (details in the Statistical Analyses section).

This study has generally revealed causal factors in heart failure such as sex, exercise angina (a type of chest pain during performing exercises), distinctive types of chest pain, and squared Cholesterol level.

The generalisability of our results is subject to certain limitations. For instance, our data set does not cover younger generations (less than 28 years old), which will cause the analyses to be biased toward older ages. Another issue that was not addressed in this study was the mortality of the patients. This might not seem arguable at first look. However, many patients with asymptomatic chest pain might have lived without any critical problems throughout their lives, and patients with other types of chest pain might have faced devastating situations. This might cause our findings to be questionable from different perspectives. Overall, our study concluded significant risk factors in heart failure. The following chapters are a detailed description of our analyses, methods, and results.

Data Collection and exploration

The data were collected from the Kaggle website (kaggle.com), an online open-source community of data scientists and machine learning practitioners. One can easily access the online version of our data through [3]. The data contain 918 subjects with 11 covariates and one binary response (Heart Failure or not). The data do not have any missing values and are ready for analysis.

Figure 3 represents the distribution of age in samples for their sexuality and heart condition. As it is obvious, the number of male samples is dramatically higher than female samples indicating whether our data set is biased or males have more heart failure than females. According to two distinctive independent studies, men have more incidents of heart failure, which is consistent with our data set [11], [12]. Another observation in Figure 3 is that the range of the age starts from 28, showing that our data set does not contain younger generations, and our analysis is not valid for younger ages.

As shown in Appendix Figure 3 (A), asymptomatic chest pain type is more frequent in men and women, and also samples with asymptomatic chest pain type are more exposed to heart failure. This observation is not surprising because patients with heart failure may show symptoms [13]. Appendix Figure 3 (B) presents that samples with exercise-induced angina[2] are more exposed to heart failure in both sex groups.

In Appendix Figure 4, we can see that the patients with heart problems have higher oldpeak[3], and as one might suspect, the heart failure might be affected by the quadratic of the oldpeak as well. Therefore, we decided to include quadratic forms of numerical variables in our modeling as well. A detailed description of the main findings, together with our conclusion, is provided in the next chapters.

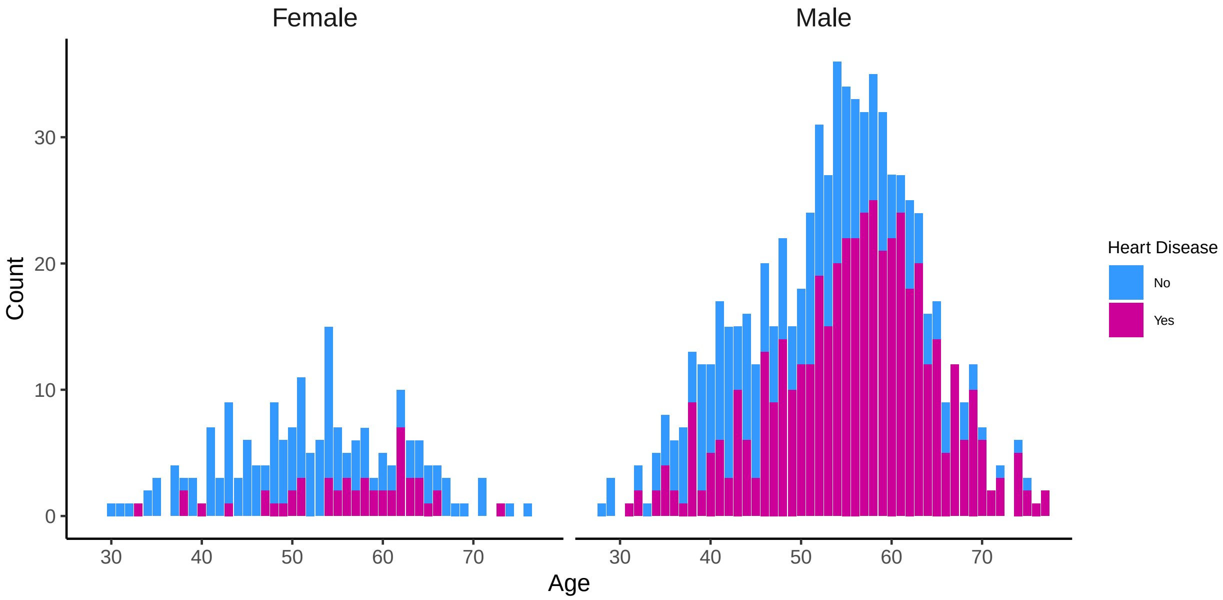


Figure 3: The distribution of the age of the samples with respect to their sex and heart condition

Statistical Analyses

Three GLM models and stepwise AIC selection

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

[2] Angina is a type of chest pain caused by reduced blood flow to the heart [14].
[3] ST depression induced by exercise relative to rest [15].

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

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

with 4 degrees of freedom. 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

Similarly, we did a Log-likelihood ratio test to the reduction, and get

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-288.2246 + 292.3055) = 8.161785$$

with 11 degrees of freedom. So the p-value is 0.699, which means the reduction is acceptable.

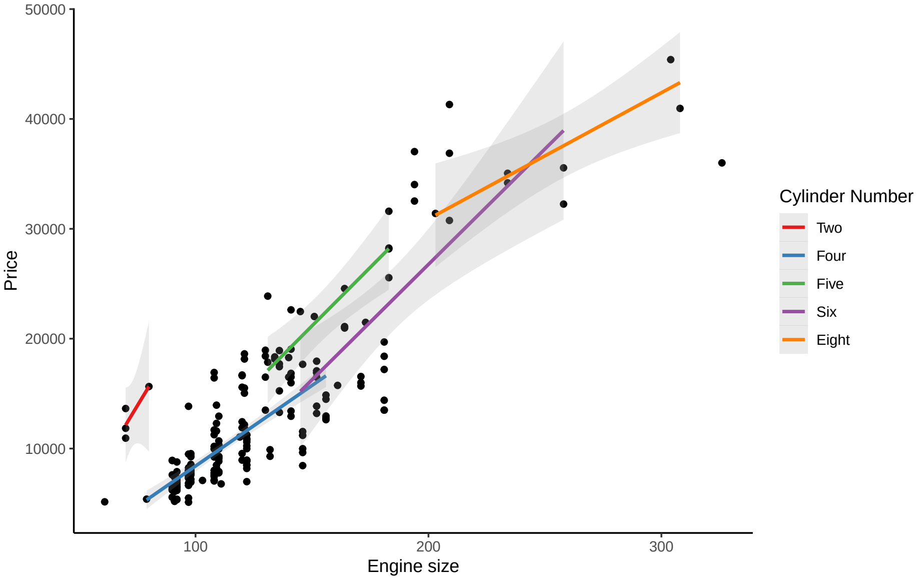
Analysis of the Three Models

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

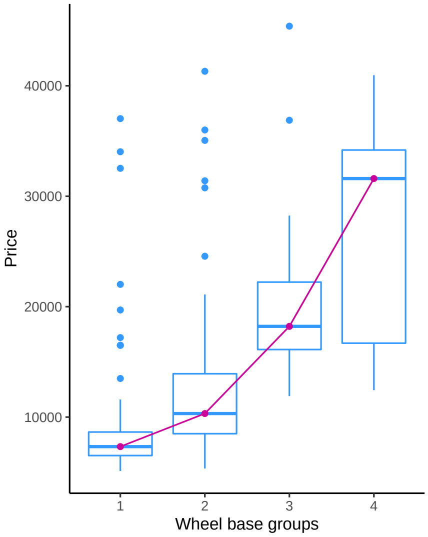
Conclusion

Appendix

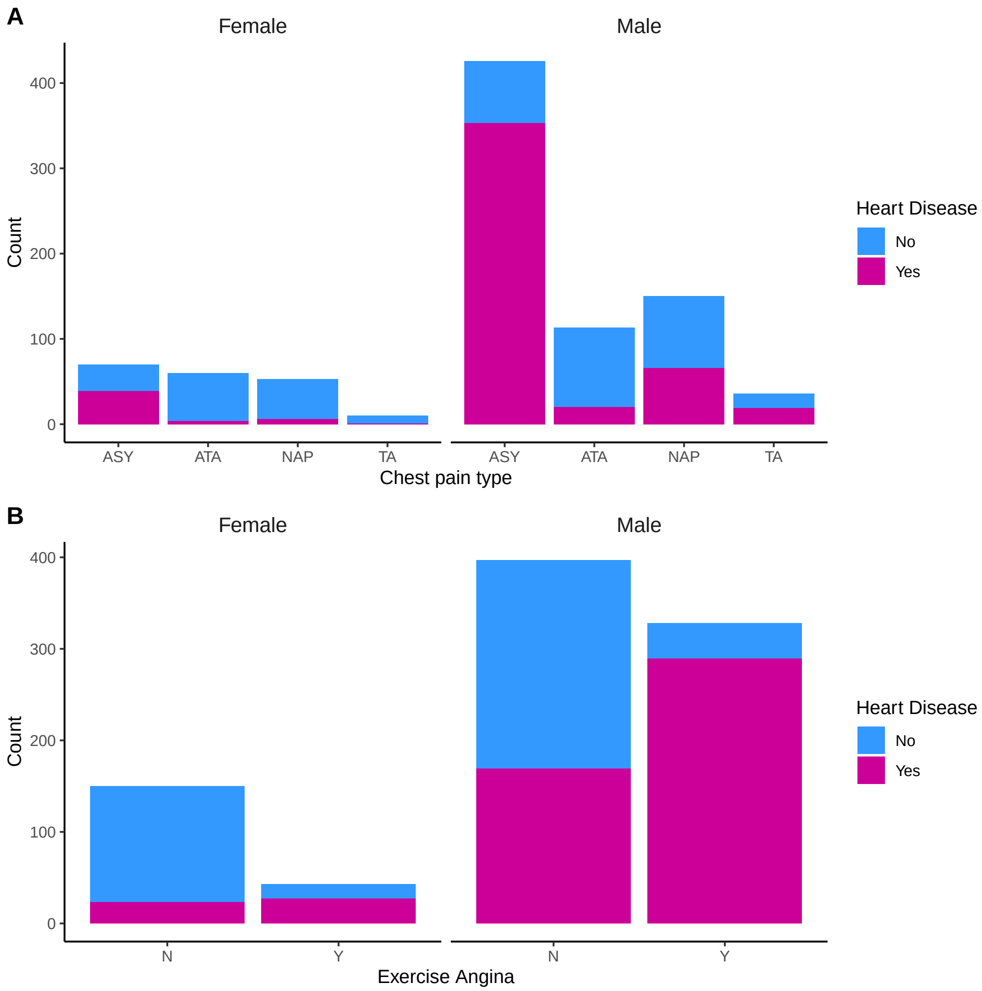
Appendix Figures



Appendix Figure 1: The correlation between engine size and price for distinctive number of Cylinders



Appendix Figure 2: The quadratic increase of price with different groups of wheelbase.



Appendix Figure 3: The number of samples with different chest pain type and exercise angina. (A) The distinctive chest pain types in different genders. TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic (B) Number of samples with exercise angina with respect to their sex. Y: Yes, with exercise angina and N: No, without exercise angina.

References

1. Dobson, A.J., Barnett, A.G.: An introduction to generalized linear models. CRC press (2018)

2. Kumar, M.: Car price prediction, https://www.kaggle.com/hellbuoy/car-price-prediction

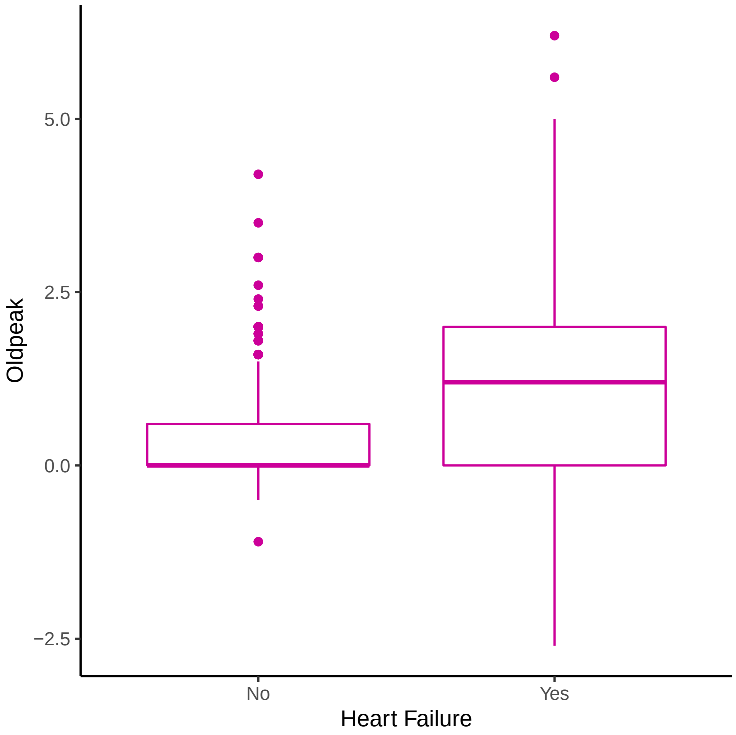
3. Palacios, F.S.: Heart failure prediction, https://www.kaggle.com/fedesoriano/heart-failure-prediction

4. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016)

5. Analysis Unit (I&A), I.T.A.I. &: The automotive industry in the united states, https://www.selectusa.gov/automotive-industry-united-states

6. Kane, M.: US: All-electric car market share expands to 2.5, https://insideevs.com/news/526699/us-electric-car-registrations-2021h1/

7. Wikipedia: Wheelbase, https://en.wikipedia.org/wiki/Wheelbase



Appendix Figure 4: The relationship between having heart disease and the oldpeak.

8. National Heart, Lung, Blood Institute, N.I. of H.(NIH).: Heart failure, https://www.nhlbi.nih.gov/health-topics/heart-failure

9. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N., others: Heart disease and stroke statistics—2020 update: A report from the american heart association. Circulation. 141, e139–e596 (2020)

10. Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., others: Heart disease and stroke statistics—2019 update: A report from the american heart association. Circulation. 139, e56–e528 (2019)

11. Strömberg, A., Mårtensson, J.: Gender differences in patients with heart failure. European Journal of Cardiovascular Nursing. 2, 7–18 (2003)

12. Mehta, P., Cowie, M.R.: Gender and heart failure: A population perspective. Heart. 92, iii14–iii18 (2006)

13. Canto, J.G., Shlipak, M.G., Rogers, W.J., Malmgren, J.A., Frederick, P.D., Lambrew, C.T., Ornato, J.P., Barron, H.V., Kiefe, C.I.: Prevalence, Clinical Characteristics, and Mortality Among Patients With Myocardial Infarction Presenting Without Chest Pain. JAMA. 283, 3223–3229 (2000). https://doi.org/10.1001/jama.283.24.3223

14. Harvard Health Publishing, H.M.S.: Angina: Symptoms, diagnosis and treatments, https://www.health.harvard.edu/heart-health/angina-symptoms-diagnosis-and-treatments

15. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS international conference on computer systems and applications. pp. 108–115. IEEE (2008)