Individual Project Final Report

Jacob Li

Business Understanding

According to Centers for Disease Control and Prevention, heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States: about 655,000 Americans die from heart disease each year—that's 1 in every 4 deaths.[1] Among all types of heart disease, Coronary Artery Disease (CAD) is the most common type, about 18.2 million adults age 20 and older have CAD in the United States.[2]

However, it is a long and winding way for doctors to make definite diagnoses for CAD. Even for the most knowledgeable doctors, the only way to make definite diagnoses on this disease is to conduct cardiac angiography (imaging) surgery. The problem is, this surgery is not suitable for everyone: even if modern medicine has made it a quite riskless surgery, it may cause postoperative complications due to bleeding; for older patients or patients with cardiac insufficiency, this surgery might even cause death. Another issue is that cardiac angiography surgery is much more expensive compared with routine medical testing, but essentially it is only a way to diagnose but not cure. In this case, some patients might find conducting this surgery unfavorable even if some symptoms have shown, because there are many other kinds of heart disease that may induce

---

[1] Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. Circulation. 2020;141(9):e139–e596.

[2] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

similar discomforts and cannot be detected from this surgery. This fact makes the cost performance of this surgery relatively low. Last but not the least, if diameter narrowing of the major vessels is less than 70%, even conducting cardiac angiography surgery might not detect CAD successfully.

Under this circumstance, the goal of this project is to build a model that could help doctors decide whether a cardiac angiography surgery is necessary according to patients' specific conditions. The logic here is that if the probability that the patient have CAD is high (model indicates a more than 50% diameter narrowing), then a cardiac angiography surgery might be more useful and deserve the risk. It can also help patients to make decisions too. If the probability of having CAD indicated by the model is low, then the patient could possibly regard whatever symptoms they have as caused by other types of heart disease and skip this surgery to save money.

Data Understanding

The dataset I used to build this model is on Kaggle. However, Kaggle is by no means the source. The original database contains 76 variables and was collected from four locations in 1988: Cleveland Clinic Foundation, Hungarian Institute of Cardiology located at Budapest, V.A. Medical Center located at Long Beach, CA and University Hospital located at Zurich, Switzerland.

Among all 76 variables, all published experiments refer to using a subset of 14 of them (including the response variable "target"). This subset is what was posted on Kaggle and used to do this project. Please see the 14 effective

variables as below:

Variable	*Type*	*Explanation*
age	Numerical	Age of individual
sex	Categorical	--Value 1=male, --Value 0=female
cp	Categorical	Chest pain type. -- Value 0: asymptomatic -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain
trestbps	Numerical	Resting blood pressure (in mm/Hg on admission to the hospital)
chol	Numerical	Serum cholestoral in mg/dl
fbs	Categorical	Fasting blood sugar >120 mg/dl: --Value 0: False --Value 1: True
restecg	Categorical	Resting electrocardiographic results: --Value 0: normal --Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) --Value 2: showing probable or definite left ventricular hypertrophy
thalach	Numerical	Maximum heart rate achieved
exang	Categorical	Exercise induced angina --Value 1 = yes; --Value 0 = no
oldpeak	Numerical	ST depression induced by exercise relative to rest
slope	Categorical	The slope of the peak exercise ST segment -- Value 0: upsloping -- Value 1: flat -- Value 2: downsloping
ca	Numerical	Number of major vessels (0-3) colored by fluoroscopy
thal	Categorical	--Value 1 = normal; --Value 2 = fixed defect; --Value 3 = reversable defect
target	Categorical	Diagnosis of heart disease -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing

In general, there are two potential issues related to this dataset. One is that some seemingly relevant factors such as smoking habit and dietary structure are

not included. We have to believe in previous experiments upon which these 14 variables were picked. Another issue is that the data was collected in 1988 from only four locations.

These issues related to this dataset requires the model to have the following assumptions:

Assumptions
1. Previous experiments were properly designed and conducted by professional researchers so that their conclusions are tenable.
2. Human being's inner system works in the same way today as it worked in 1988.
3. Living environment doesn't affect the chance of having CAD.

There seems to have selection bias in this dataset since people who took these tests and got recorded in hospitals' databases were more likely to have CAD. However, since this model is primarily built to discuss whether conducting cardiac angiography surgery is necessary for people who already turn to hospitals, selection bias is not a problem here.

Data Preparation

After further probing into the dataset, I noticed that there were still some flaws regarding the interpretation of variables. There are supposed to have 3 distinct values in thal (which are 1 to 3 as presented in the table) and 4 distinct values in ca (which are 0 to 3). However, in the raw dataset, there were some value 0 in thal and value 4 in ca. These additional categories couldn't be interpreted.

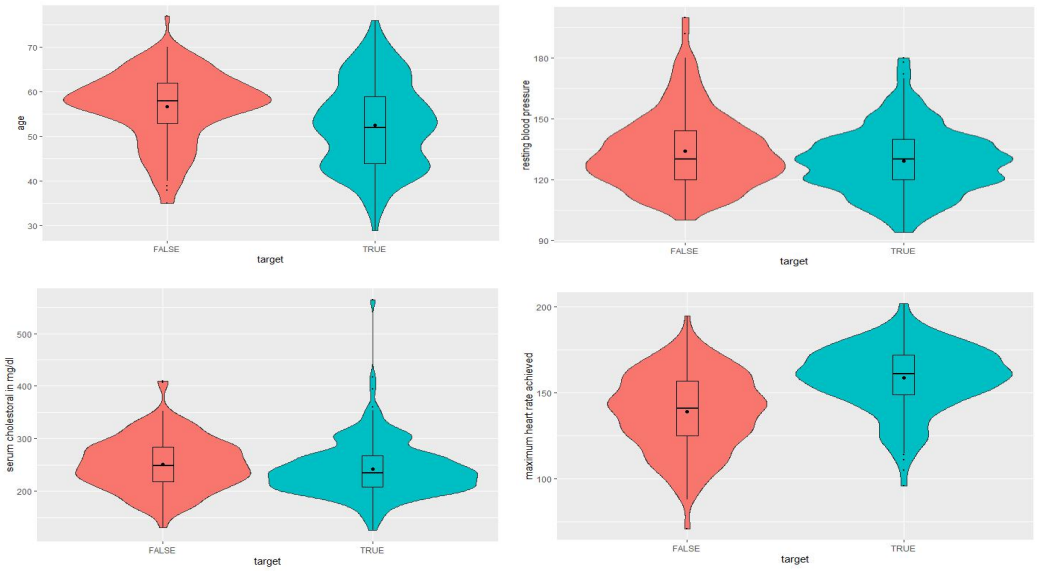
Fortunately EDA showed that there were only 25 rows of them in total, in this case I just deleted these rows. Up to this point, there were 1000 rows in this dataset, without any NA values.

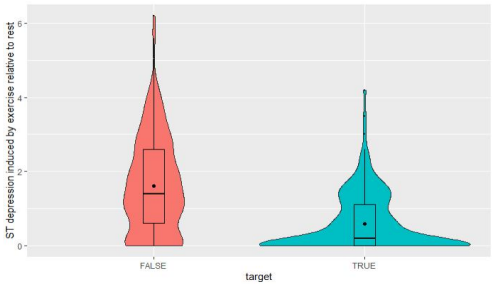
To further prepare the data for modeling, I looked into the overall rate of target being value 1, and the result is approximately 0.51, which means the dataset is balanced and doesn't need bootstrapping to generate more negative "labels". Moreover, I changed categorical variables into factors and named target equals one "TRUE" and target equals zero "FALSE" to improve plotting efficiency.

Modeling and Evaluation

Exploratory Data Analysis:

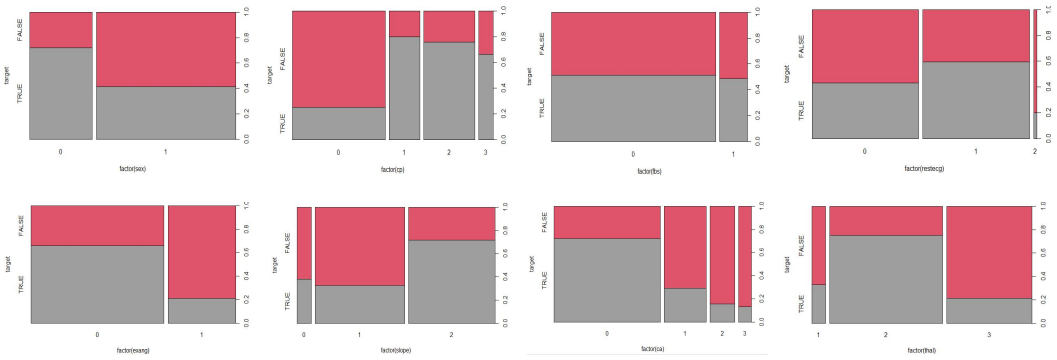
The violin plots for numeric variables against the binary target variable with red as FALSE and blue as TRUE are as follows: (in the sequence of age, trestbps, chol, thalache and oldpeak, from left to right by row)





We can see that among age, thalache and oldpeak, each of them has very distinct distributions in terms of different value of target. By comparison, trestbps and chol only show very small differences between each group of target. However, considering that in trestbps when target is TRUE there are more volatility around the mean, and in chol there are a lot of interesting outliers when target is TRUE, I decided to keep these two variables.

Similarly, I made the following plots for each categorical variable, with red area representing the percentage of target is FALSE and grey area representing the percentage of target is TRUE: (in the sequence of sex, cp, fbs, restecg, exang, slope, ca and thal from left to right by row)



Despite these plots are tiny in size, we can still clearly see that compared with the “performance” of other variables, the third variable fbs seems to be indifferent with our reponse variable; so I took a closer look at its p-value and the result was around 0.63, which proved my thought that independence couldn’t be rejected.

In conclusion, I only sifted out one varibale fbs as noice according to the EDA. Since the response variable is a binary, it's hard to find interactions simply using plots. In this case, a Lasso Selection was also conducted.

Lasso Regularization and Selection

The following picture shows the result of Lasso selection. (The code is adapted from material in ScriptClass8.R that professor posted on Canvas.)

```
 [1]  11  13  37  44  53  54  67  74  81  85 108 116 150 167 169 171 179 185 186 198
 [1] "thalach"          "oldpeak"          "age:ca3"          "sex1:chol"
 [5] "sex1:ca1"         "sex1:ca2"         "cp1:restecg1"     "cp2:thalach"
 [9] "cp3:oldpeak"      "cp1:slope2"       "trestbps:exang1"  "trestbps:thal3"
[13] "restecg1:slope2" "thalach:ca2"      "thalach:thal2"    "exang1:oldpeak"
[17] "oldpeak:slope1"  "oldpeak:thal3"    "slope1:ca1"       "ca3:thal2"
```

Lasso only picked out two variables as good predictors, which was surprising to me since the plots in EDA showed quite different conclusion. However, it certainly provided a lot of information regarding hidden interactions. In this case, to be more rigorous, I designed the 10-fold cross validation in the following way.

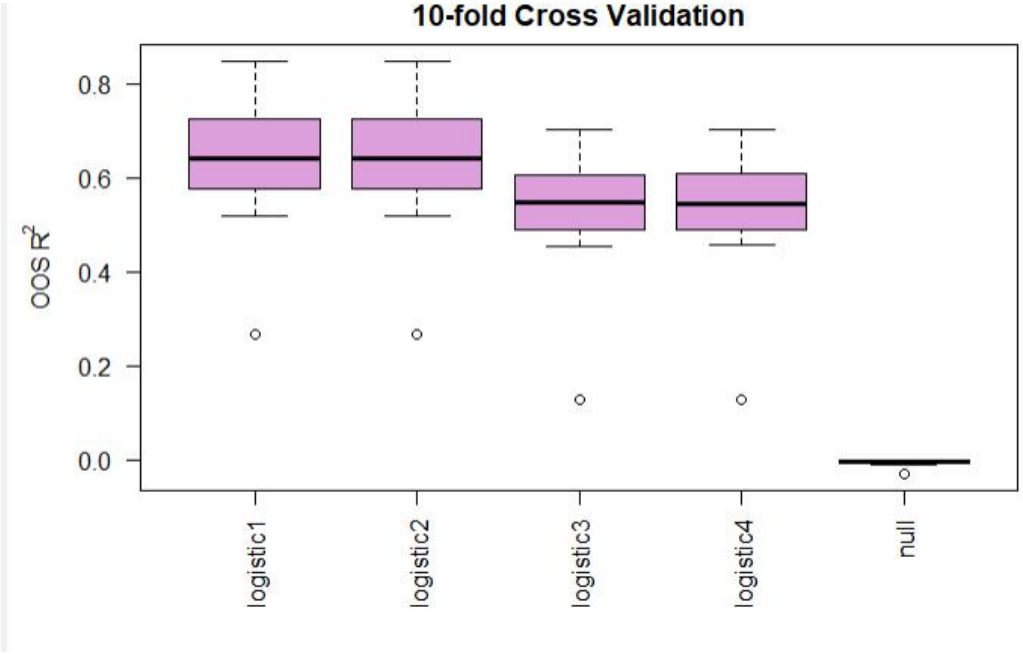
10-fold Cross Validation – Models and Evaluation

First, I tested four logistic models' out-of-sample R-squared using 10-fold cross validation (the code is also adapted from material in ScriptClass8.R that professor posted on Canvas). The specific models are shown below in the table:

Type & Name	Variables Fitted	Explanation
Logistic1	All interactions identified by Lasso and all variables in the original dataset except fbs	This is based on the combination of Lasso's choice and my understanding
Logistic2	All Lasso's selection, including those that are not interaction	This is purely based on Lasso's choice
Logistic3	All variables in the original dataset except fbs	This is purely based on my understanding

Logistic4	All variables in the original dataset	This is the “control group”

When also compared with the null model which predicts all potential patients equally has the same possibility of having diameter narrowing more than 50%, the performance of these four models is shown below:




```
  logistic1   logistic2   logistic3   logistic4        null
0.62391737  0.62391737  0.52226154  0.52345736 -0.00569488
```

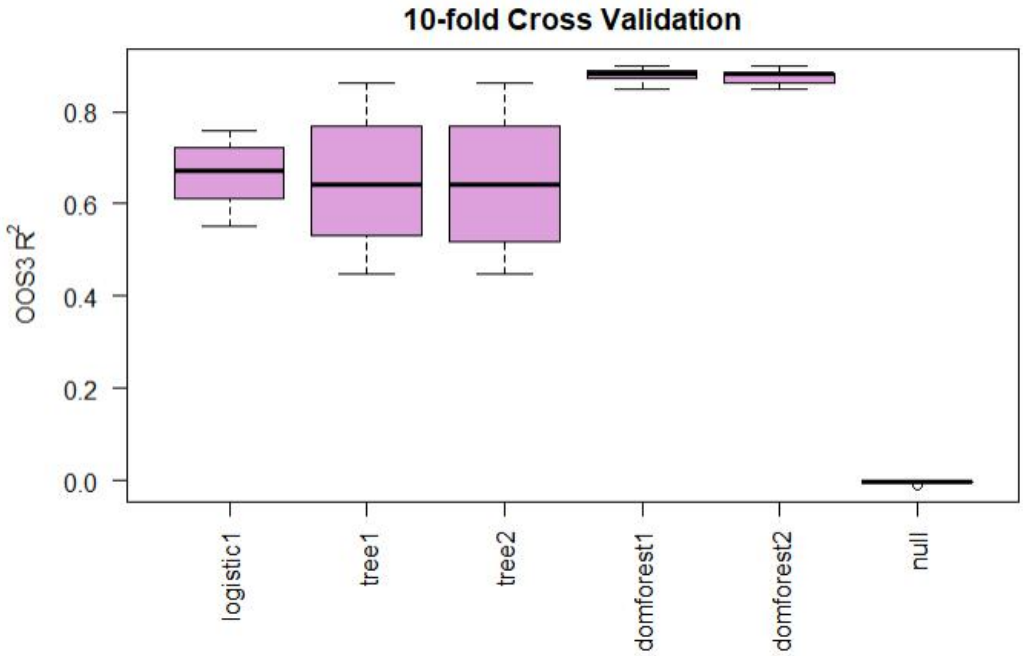
We can see that the OOS R-squared of logistic1 and logistic2 have exactly the same mean, which infers that probably all predictor variables in the original dataset doesn’t matter for logistic models except for the two that Lasso identified: thalach and oldpeak. The numbers below the plot also indicates that excluding fbs is not a good choice, since the mean of R-squared slightly decreases when interquartile range remains unchanged by pure observation.

The aforementioned process is followed by another similar one. I picked out the model that performed the best, which is logistic1 or logistic2 in this case, and compared its OOS R-squared with those of classification trees and Random Forest using 10-fold cross validation again. The specific models are shown below

in the table.

Type & Name	Variables Fitted	Explanation
Logistic1	All interactions identified by Lasso and all variables in the original dataset except fbs	This is based on the combination of Lasso's choice and my understanding
Tree1	All variables in the original dataset except fbs	This is a tree model reflecting my understanding
Tree2	All variables in the original dataset	This is built to be a comparison with Tree1
Randomforest1	All variables in the original dataset except fbs	This is a random forest reflecting my understanding (arguments: ntree=500,nodesize=5,mtry=4)
Randomforest2	All variables in the original dataset	This is built to be a comparison with Randomforest1 (arguments have the same values)

The performance of these 5 models is shown below:




```
  logistic1         tree1         tree2 randomforest1 randomforest2          null
0.663871357   0.642229142   0.641026321   0.876962224   0.875007452 -0.003062885
```

We can see from the plot that logistic1 is no longer the winner. With a high out-of-sample R-squared and short interquartile range, randomforest1 is the

best-performing model compared to all the other models. Also, tree1 and randomforest1 both performed slightly better than their deliberately built comparison, which means that excluding fbs is a better choice for classification tree and Random Forest model.

In conclusion, randomforest1 should be the model used to predict the target variable in this project, because not only its OOS R-squared is very high, its robustness is also satisfying. With that being said, one potential direction for further improvement could be tweaking the values for the arguments, such as changing the input for mtry and nodesize to other numbers to see if the result becomes even better.

In addition to the evaluation of models, another evaluation topic is the ROI of the project. However, considering that the eccense of this project is to improve the efficiency of a critical medical process, I do think it should be charitable and open-sourced; in this case there is no consideration regarding ROI.

Deployment

The audience of this model should be hospitals. With hospital's approval, their vasculocardiology department may install this model onto their internal system. Doctors may input their patient's information into the model, and the more testing a certain patient has taken, the more information doctors are able to input into the model, so that the precision of the model can also increase. At last, doctors and patients can look into the results together and have discussions about whether cardiac angiography surgery is necessary. Despite this model is useful in

predicting CAD, I do not expect that people can use it outside the hospital. One reason is that many predictor variables could only be gleaned from medical testing, the other reason is that without professional knowledge potential patients could be misled by the model and mistakenly underestimate his or her state of illness, which could lead to terrible consequences.

When it comes to ethical considerations, this model entails confidential information of patients. However, considering that only hospitals would have access to install this model, I believe this is not a big issue as long as the distribution of this model is controlled carefully, since qualified hospitals should already be sophisticated in this field.