BANA COMPETITION – HEART ATTACH ANALYSIS

Data analytics approach

Student name: JIZE CHEN
Student number: 46278366

Agenda

1. Explanation of heart attack
2. Data Exploration
3. Model Development
4. Interpretation
5. Limitation of model

Explanation of heart attack

What is heart attach

- A heart attack occurs when the flow of blood to the heart is severely reduced or blocked
- Symptoms
  - Chest pain
  - Pain or discomfort that spreads to the shoulder, arm, back, neck, jaw, teeth or sometimes the upper belly
  - Cold sweat
  - Fatigue
  - Heartburn or indigestion
  - Lightheadedness or sudden dizziness
  - Nausea
  - Shortness of breath

How it affect your health (complication)

- Irregular or atypical heart rhythms
- Cardiogenic shock
- Heart failure
- Inflammation of the saclike tissue surrounding the heart
- Cardiac arrest

Data quality assessment

The following Key issues might cause the incorrect and inefficient data analysis

- Accuracy: correct values
- Completeness: Data fields with values
- Consistency: Values free from contradiction
- Currency: Value up to date
- Relevancy: Data items with value meta-data
- Validity: Data containing allowable values
- Uniqueness: Records that are duplicated

Accuracy	• **Based on given data explanation, caa equal to 4 doesn't make sense, so caa==4 is removed** • **Assume thall = 0 means patient doesn't has heart defeat**
Completeness	No incomplete data found
Consistency	All values are free from contradiction
Currency	Assume all data are up to date
Relevancy	All variables are relevant
Validity	Change data type of all categorical variables from decimal to factor so R can correctly examine the data
Uniqueness	No duplicated data found

Data Exploration

Analysis the input variables of patients and determine which variables are significant factor of causing heart attack

Input variables

- age
- sex
- cp
- trtbps
- chol
- fbs
- restecg
- thalachh
- exng
- oldpeak
- slp
- caa
- thall

Output variables

- outcome

Model development

Generalized linear model

- Generate logistic regression model with glm in R
- Confusion matrix and its training error rate is shown on the right. In general, the glm model has low training error rate which is less than 15%

```
> table(glm.pred, output)
        output
glm.pred   0   1
       0 109  18
       1  29 147
> vcp = (cm[1] + cm[4])/dim(BANA_Comp_Data)[1]
> vcp
[1] 0.8590604
>
```

Interpretation

Significant factor (p test)

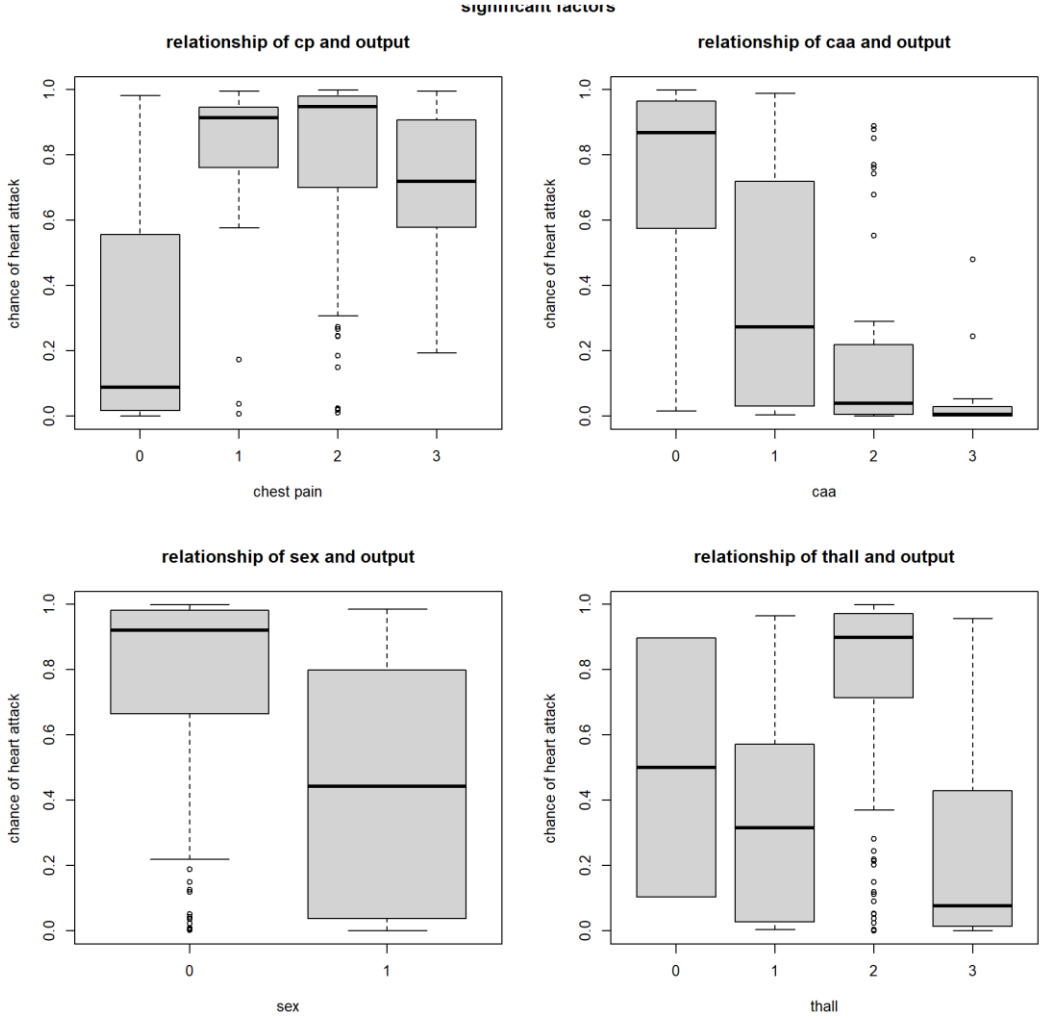
- Based on the P test on the right, most significant factors include (order by significant level)
  - cp (chest pain type)
  - caa (number of major vessels (0-3) colored by fluoroscopy)
  - sex
  - thall (heart defeat type)

```
Call:
glm(formula = output ~ ., family = binomial, data = BANA_Comp_Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8090  -0.3051   0.1342   0.4868   2.8167

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.744843   3.491652   0.213 0.831077
age          0.014546   0.025052   0.581 0.561471
sex1        -1.711361   0.542918  -3.152 0.001621 **
cp1          0.724104   0.568084   1.275 0.202436
cp2          1.833258   0.505363   3.628 0.000286 ***
cp3          2.076184   0.671459   3.092 0.001988 **
trtbps      -0.022471   0.011381  -1.974 0.048338 *
chol        -0.004228   0.004094  -1.033 0.301734
fbs1         0.554204   0.609413   0.909 0.363135
restecg1     0.398112   0.386777   1.029 0.303336
restecg2    -0.512356   2.494852  -0.205 0.837287
thalachh     0.017334   0.011173   1.551 0.120818
exng1       -0.791505   0.444412  -1.781 0.074910 .
oldpeak     -0.370507   0.232120  -1.596 0.110447
slp1        -0.698495   0.848926  -0.823 0.410622
slp2         0.600288   0.919960   0.653 0.514069
caa         -1.369208   0.286616  -4.777 1.78e-06 ***
thall1       2.266197   2.462440   0.920 0.357413
thall2       2.196570   2.374903   0.925 0.355013
thall3       0.791202   2.383423   0.332 0.739919
```

Interpretation

Who has higher chance of getting heart attack

- On the right is the graph of significant factors of heart attack which include cp, caa, sex and thall
- Patient with following characteristic has higher chance of heart attack
  - patient suffers from atypical angina, non-typical angina and asymptomatic
  - Patient who has 0 major vessels colored by fluoroscopy
  - Patient who is male
  - Patient who is fixed defeat



Limitation and possible solutions of model

Limitations

1. Because whole data set was used for training and testing, the variance of this model is low. In other word, the prediction power of this model on new data will be relatively low
2. Most of the variables doesn't explain much of the output of the model which might also cause the model of overfitting
3. Other models might outperformed logistic regression model in this data set

Possible solutions

1. Use cross validation approaches like leave 1 out validation, k fold validation to improve the model on testing data set
2. Use dimension reduction method like principal component analysis or partial least square to mitigate overfitting
3. Construct models using methods like linear probability model, polynomial model, k-nearest neighbour model and compute their training error and testing error to determine which model has the best prediction power for this data set

Reference

- Heart attack - Symptoms and causes. (2022). Retrieved 4 October 2022, from https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106
- Heart attack - Symptoms and causes. (2022). Retrieved 4 October 2022, from https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106#:~:text=A%20sudden%20change%20in%20the,cardiac%20death)%20without%20immediate%20treatment.