THE STUDY OF UCI HEART DISEASE CLEVELAND

Executive Summary

In order to complete a final project to reflect our learnings in this term's MIE 1413 course and also to study the UCI Heart Disease Cleveland dataset, we prepared this final report to provide our analysis results and findings. The purpose of this project is to figure out the relationship between heart disease and some explanatory variables among a subset of 14 variables from 76 attributes in the UCI Heart Disease Cleveland dataset.

In this report, by studying the dataset and investigate the descriptive statistics of the dataset, it is known that the dependent variable is condition, which is indicate by 0 and 1, with 13 predictor variables where 5 of them are numerical and 8 of them are categorical, 297 samples are involved in the data. After studying for the multicollinearity, the automated model selection model is first used based on the logistic regression. After the automated model selection, it is found that every term of the model selected by stepwise selection has very large p-value. Therefore, main effect model is used instead as the starting model of manual model selection. By performing deviance test to compare models repeatedly, the final model is then determined. The model diagnostic is then performed to make sure that the model satisfies all the assumptions, and the goodness of fit has been tested. In addition, conclusion is provided along with limitations of the study due to the small size dataset and samples are all in Cleveland. Future developments are also provided which might be useful for further improve the study.

TABLE OF CONTENTS

LIST OF TABLES & PLOTS

1.0 Introduction

Heart disease is the leading cause of death globally, which takes away approximately 17.9 million lives per year (World Health Organization, 2022). In order to study what factors are relevant to imply the existence of heart disease, the observational data: UCI Heart Disease Cleveland obtained from the UCI Repository is used in the study of this report. Here is the link to the dataset used in this study on Kaggle: https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci. This dataset contains 76 attributes and since all published experiments prefer to only study a subset of 14 of them, in this study the subset of 14 of them is used as well. The dataset used in this study includes 297 samples and data cleaning has been performed. The goal of the study is to explore relationship between heart disease and some explanatory variables.

The dependent variable in the dataset is condition. It is represented by the indicator variable 0 and 1, where 0 represents the person do not have heart disease and 1 represents the person have heart disease. There are 13 predictor variables with both categorical and numerical, details of them are listed as below:

Numerical (5):

- age: age in years
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- thalach: maximum heart rate achieved
- oldpeak: ST depression induced by exercise relative to rest

Categorical (8):

- sex: sex (1 = male; 0 = female)
- cp: chest pain type
  -- Value 0: typical angina
  -- Value 1: atypical angina
  -- Value 2: non-anginal pain
  -- Value 3: asymptomatic
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: 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 by Estes' criteria

- exang: exercise induced angina (1 = yes; 0 = no)
- slope: the slope of the peak exercise ST segment

  -- Value 0: upsloping

  -- Value 1: flat

  -- Value 2: downsloping
- ca: number of major vessels (0-3) colored by fluoroscopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

  and the label

The study is based on the dependent variable and predictor variables listed above.

2.0 Descriptive Statistics

2.1 Summary Statistics

From the summary statistics listed below, for categorical variables it presents different levels of categorical variables and how the data is distributed among these levels, and for numerical variables, the extreme values of the data could be known by looking at the minimum value and maximum value and the average of the value along with the 1st quantile, the median and the 3rd quantile are also presented in the summary statistics. Among the 297 samples involved in this study, 160 sample do not have heart disease while 137 samples do have heart disease.

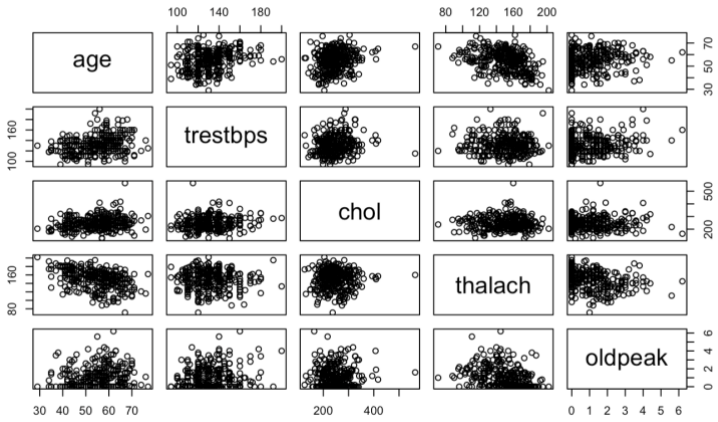
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
Min. :29.00	0: 96	0: 23	Min. : 94.0	Min. :126.0	0:254	0:147	Min. : 71.0	0:200
1st Qu.:48.00	1:201	1: 49	1st Qu.:120.0	1st Qu.:211.0	1: 43	1: 4	1st Qu.:133.0	1: 97
Median :56.00		2: 83	Median :130.0	Median :243.0		2:146	Median :153.0	
Mean :54.54		3:142	Mean :131.7	Mean :247.4			Mean :149.6	
3rd Qu.:61.00			3rd Qu.:140.0	3rd Qu.:276.0			3rd Qu.:166.0	
Max. :77.00			Max. :200.0	Max. :564.0			Max. :202.0	

oldpeak	slope	ca	thal	condition
Min. :0.000	0:139	0:174	0:164	0:160
1st Qu.:0.000	1:137	1: 65	1: 18	1:137
Median :0.800	2: 21	2: 38	2:115	
Mean :1.056		3: 20		
3rd Qu.:1.600				
Max. :6.200				

Table 1: Summary Statistics

2.2 Pair Plot

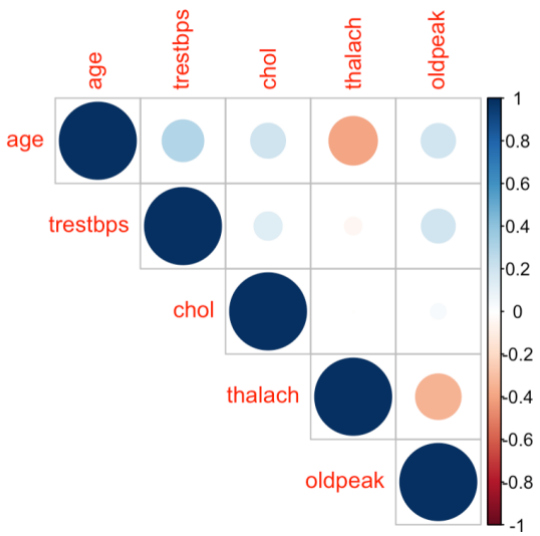
From the pair plot of all the numerical variables, there is no evidence that there exist any strong colinear relationship among these variables, so there is no evidence that any variable will cause concern of multicollinearity and no variable will be removed at this step.



Plot 1: Pair Plot

2.3 Correlation Plot

Below is the correlation plot of the continuous variables. There are some correlations between variables, especially between oldpeak and thalach, age and trestbps. Therefore, a further step is done to test the variance inflation factor (VIF).



Plot 2: Correlation Plot

2.4 Variance Inflation Factor

The VIF explains the amount of variance of a predictor variable is influenced by its interaction or correlation with other predictor variables, and it measures the amount of multicollinearity where high multicollinearity should be avoided from the model (The Investopedia Team, 2022). By checking the VIF, it has been found that the VIF of age is 1.35, the VIF of trestbps is 1.14. the VIF of chol is 1.06, the VIF of thalach is 1.33 and the VIF of oldpeak is 1.18. By looking at these VIF, they are all relatively small, which do not raise concern of multicollinearity and thus could all be kept in the data.

With the above analysis of the descriptive statistics, it is possible to proceed to building the models.

3.0 Model Selection

Since the dependent variable condition has a binary response of 0 or 1, and both numerical variables and categorical independent variables are included in the study, logistic regression is used, and model selection are based on the binary logistic model.

3.1 Automated Model Selection

In order to find a good model, the automated model selection process is used as the starting point. Three different methods are used, which are forward selection method that starts from the model with intercept only, backward elimination method that starts from the full model, and stepwise selection method that starts from the main effect model. The system runtime of backward elimination method is much longer than that of stepwise selection method and forward selection method, and when the dataset is very large, it would raise some concerns towards the much longer runtime. Therefore, it would be better to proceed with comparing the model generated by stepwise selection method and forward selection method instead, which has runtime of 2.21 seconds and 0.34 seconds respectively. By comparing the AIC of these two models, model selected by forward selection has 36 degrees of freedom with AIC equals to 206.41, while stepwise model has 61 degrees of freedom with AIC equals to 150.40, therefore, since the model selected by stepwise selection method has much lower AIC, this model is selected as the result of automated selection method. However, from the summary table of this model, it shows that when interaction terms

have been added to the model, all p-values are very large and its corresponding terms are very not significant, even if the model is selected by automated model selection. Therefore, it is possible that adding these interactions would lead to large p-value in this model because of not so many samples contained in the model, so it is preferring to stay with the main effect model at this stage.

3.2 Manual Model Selection

From the summary output of the main effect model listed below, it is necessary to find out the variable that has the largest p-value, which is resting electrocardiographic results, and then remove this explanatory variable from the model, which result in a reduced model.

```
glm(formula = condition ~ age + sex + cp + trestbps + chol +
    fbs + restecg + thalach + exang + oldpeak + slope + ca +
    thal, family = binomial(link = "logit"), data = heart_disease)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0490  -0.4847  -0.1213   0.3039   2.9086

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.253978   2.960399  -2.113 0.034640 *
age         -0.023508   0.025122  -0.936 0.349402
sex1         1.670152   0.552486   3.023 0.002503 **
cp1          1.448396   0.809136   1.790 0.073446 .
cp2          0.393353   0.700338   0.562 0.574347
cp3          2.373287   0.709094   3.347 0.000817 ***
trestbps     0.027720   0.011748   2.359 0.018300 *
chol         0.004445   0.004091   1.087 0.277253
fbs1        -0.574079   0.592539  -0.969 0.332622
restecg1     1.000887   2.638393   0.379 0.704424
restecg2     0.486408   0.396327   1.227 0.219713
thalach     -0.019695   0.011717  -1.681 0.092781 .
exang1       0.653306   0.447445   1.460 0.144267
oldpeak      0.390679   0.239173   1.633 0.102373
slope1       1.302289   0.486197   2.679 0.007395 **
slope2       0.606760   0.939324   0.646 0.518309
ca1          2.237444   0.514770   4.346 1.38e-05 ***
ca2          3.271852   0.785123   4.167 3.08e-05 ***
ca3          2.188715   0.928644   2.357 0.018428 *
thal1       -0.168439   0.810310  -0.208 0.835331
thal2        1.433319   0.440567   3.253 0.001141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2: Summary Table of the Main Effect Model

In order to compare which model fits the data better, a deviance test has been conducted with the null hypothesis that the reduced model provides a better fit whereas the alternative hypothesis states that the main effect model provides a better fit. The p-value of the test is 0.45 which is much larger than 0.05, so fail to reject the null hypothesis, and can conclude that the reduced model provides a better fit at 0.05

significance level. Then, the steps of moving the explanatory variable that corresponds to the largest p-value in this model to build a new reduced model then followed by a deviance test to compare the performance of the two models has been repeated until a final model has been achieved.

3.3 Final Model

Finally, the final model has been achieved, and below is the summary output of the final model. There are 7 predictor variables included in the final model which includes two numerical variables that are resting blood pressure and ST depression induced by exercise relative to rest, and 5 categorical variables that are sex, chest pain type, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy and level of defect.

```
glm(formula = condition ~ sex + cp + trestbps + oldpeak + slope +
    ca + thal, family = binomial(link = "logit"), data = heart_disease)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8068  -0.5022  -0.1270   0.3829   2.9998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.86579    1.89438  -4.680 2.87e-06 ***
sex1         1.39946    0.49810   2.810 0.004960 **
cp1          1.45775    0.79649   1.830 0.067219 .
cp2          0.31192    0.69677   0.448 0.654392
cp3          2.75456    0.68831   4.002 6.28e-05 ***
trestbps     0.02469    0.01071   2.305 0.021182 *
oldpeak      0.49800    0.22399   2.223 0.026194 *
slope1       1.53979    0.45813   3.361 0.000776 ***
slope2       0.65270    0.86245   0.757 0.449172
ca1          2.30806    0.48468   4.762 1.92e-06 ***
ca2          2.83132    0.71991   3.933 8.39e-05 ***
ca3          2.19538    0.90437   2.428 0.015203 *
thal1       -0.08340    0.74388  -0.112 0.910736
thal2        1.52373    0.42081   3.621 0.000294 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3: Summary Table of the Final Model

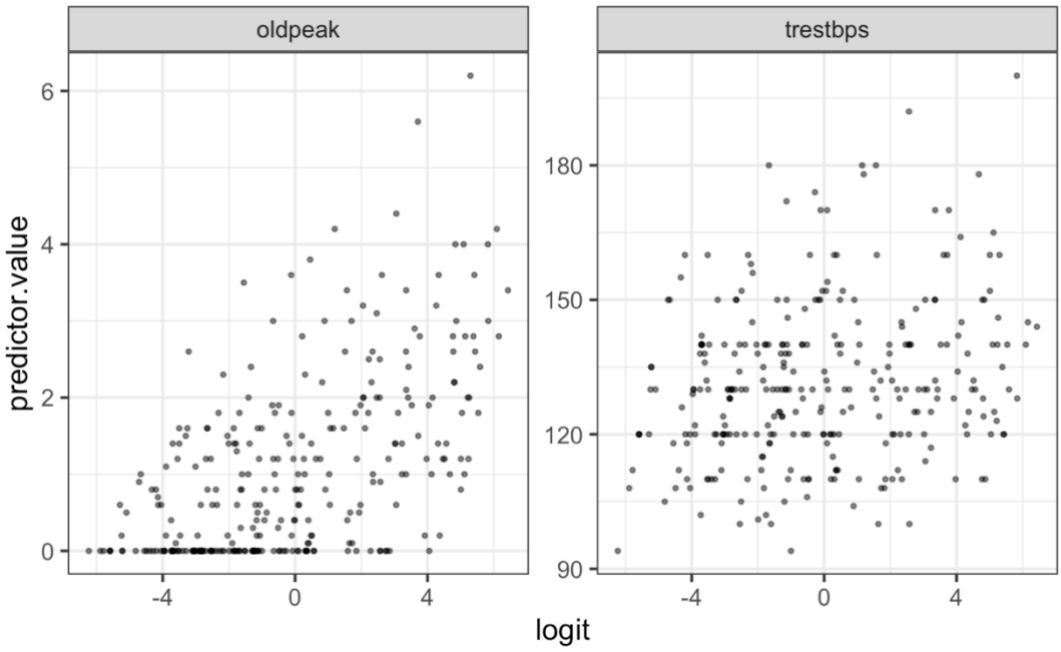
4.0 Model Diagnostic

After selecting the final model, a Regression diagnostic is performed. It is used to evaluate the model assumptions and investigate the goodness of fits.

Firstly, check has been performed to check if the model is consistent with the logistic model's assumption. There are three main assumptions of the logistic model: linearity

assumption, independence assumption and multicollinear assumption. The first assumption is that the numerical predictor variables will have a linear relationship with the outcome variable's logit. This linearity assumption can be checked through the scatter plot between the logit outcomes and each predictor. (Logistic Regression Assumptions and Diagnostics in R - Articles - STHDA, n.d.)

In the model, the numerical predictor variables are ST depression induced by exercise relative to rest and resting blood pressure. The y-axis of the scatter plot on figure 4.1 is the predictor value, and the x-axis is the logit of the outcome. The scatter plots below (plot 3) shows that both variables are quite linearly associated with the heart disease outcome in logit scale since the points are randomly distributed.



Plot 3: Scatter Plot Between the Predictor Value of Variable oldpeak (left) and trestbps (right) and the Logit Values

The second assumption is that the observations in logistic regression should be independent with each other. It is not allowed that the observations come from repeated measurements or matched data. Since the researchers do not record the same individuals at different times, this assumption is met.

The third assumption is that multicollinearity should not appear among the predictor variables in logistic regression. In other words, the correlation between predictor variables should not be too high. In part 4, rather than checking the VIF as did in part 2, Generalized VIF is used to check this assumption. The variable contains more than one degree of freedom, so VIF cannot be applied to check multicollinearity. The third column of the GVIF table (table 4.2) should be paid attention to, it indicates the GIF to the power of one over two times degrees of freedom ($GVIF^{\frac{1}{2DF}}$). The degree of freedom is the number of coefficients in the subsets. This reduces the GVIF to a linear measure. It is equivalent to taking the square root of the usual VIF (Buteikis, n.d.). It shows that all the number on the third column is smaller than 2, which indicates that multicollinearity does not exist in this model.

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.439324	1	1.199719
cp	1.605205	3	1.082069
trestbps	1.128841	1	1.062469
oldpeak	1.519753	1	1.232783
slope	1.704172	2	1.142558
ca	1.436789	3	1.062263
thal	1.354106	2	1.078731

Table 4: GVIF Table

Secondly, deviance is used to check the overall goodness of fit for the model. The null hypothesis (H0) is that the model fits, while the alternative hypothesis (H1) is that the model does not fit. If a deviance is much higher than the number of observations minus the number of parameters in the model, then the model is a poor fit to the data. In general, the smaller the deviance, the better the fit of the model.

The Residual Deviance for the model in question is 192.6 according to the deviance output. The deviance is equal to -2LL (-2 multiply the model of interest's log-likelihood). Then, pass the residual deviance, 192.6 along with the model degrees of freedom, 283 to χ^2 to determine whether there is strong evidence to reject the null hypothesis. Recall that the null hypothesis is that the model is appropriate (Triveri,

2017). Since the p-value is larger than significant level 0.05, there is no evidence to reject the null hypothesis, the model fits.

5.0 Result

The model used here is logistic regression, where the response (proportion of having heart disease) is linked to a linear combination of covariates with a logit link.

For the i_{th} patient, where μ_i is the proportion of having heart disease.

$$\log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

$$Y_i \sim Binomial(N_i, \mu_i)$$

There are seven covariates:

- sex (gender, categorical with two levels: 1 = male, 0 = female)
- cp (chest pain type, categorical with four levels: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
- trestbps (resting blood pressure (in mm Hg on admission to the hospital), numerical)
- oldpeak (ST depression induced by exercise relative to rest, numerical)
- slope (the slope of the peak exercise ST segment, categorical with three levels: 0 = upsloping, 1 = flat, 2 = downsloping)
- ca (number of major vessels (0-3) colored by fluoroscopy, categorical)
- thal (whether the observation have detective, categorical with three levels: 0 = normal; 1 = fixed defect; 2 = reversable defect)

	est	2.5	97.5
(Intercept)	0.000	0.000	0.006
sex1	4.053	1.527	10.759
cp1	4.296	0.902	20.468
cp2	1.366	0.349	5.352
cp3	15.714	4.078	60.559
trestbps	1.025	1.004	1.047
oldpeak	1.645	1.061	2.552
slope1	4.664	1.900	11.447
slope2	1.921	0.354	10.413
ca1	10.055	3.889	25.998
ca2	16.968	4.138	69.569
ca3	8.983	1.526	52.874
thal1	0.920	0.214	3.953
thal2	4.589	2.012	10.470

Table 5: The Odds and CI for Baseline for All Variables

The coefficients of the summary table represent how odds ratio increase or decrease between groups. The second and the third columns represents the 95% confidence interval of each variable. Note that if the confidence interval involves 1, the variable will not have significant influence with the proportion of having heart disease.

According to the summary table, the coefficient of sex is above 1 and its confidence interval does not contain 1, thus it is a significant variable. The odds for male patients to have heart disease is 4.05 times of female.

For the variable chest pain, the odds of the patient with asymptomatic chest pain suffering from heart disease is 15.71 times that of the patient who has typical angina, which is different from what we imagine. However, the patients with atypical angina and non-anginal pain do not significantly differ in the odds of having heart disease from those with typical angina. The reason is that the confidence interval of cp1(atypical angina) and cp2(non-anginal pain) all contains 1.

Similarly, for the variable slope, the odds of a patient with a flat slope of the peak exercise ST-segment having heart disease are 4.664 times that of the patient with upsloping. Yet, the confidence interval of slope2 contains 1, indicating there is no significant difference between patients with upsloping and downsloping of the peak

exercise ST segment on the odds of having heart disease. Additional, for the variable thal, the odds of a patient with a reversible defect suffering from heart disease is 4.59 times that of a patient who has a normal defect. But the patients with fixed defects do not significantly differ from patients with normal defects in the odds of heart disease.

In contrast, the patient with different levels of ca all has a significant difference. The confidence interval of ca1, ca2 and ca3 all exclude 1. The patient with zero major vessels colored by fluoroscopy has the lowest odds to suffer from heart disease, while the patient with two major vessels colored has the highest odds. The odds of having heart disease among patients with one, two, and three major vessels colored by fluoroscopy are 10.1%, 17.0% and 9.0% of the patient who does not have any major vessels colored by fluoroscopy, respectively.

In addition, both resting blood pressure and ST depression induced by exercise relative to rest are significant variables in this model because both of their confidence interval exclude 1. The odds of patient have heart disease increase by 1.02 for each additional unit of resting blood pressure while the odds increase by 1.65 for each additional unit of ST depression induced by exercise relative to rest.

6.0 Conclusion & Future Developments

6.1 Conclusion

In this research project, 297 observations are used from UCI Heart Disease Cleveland to find the factors relevant to imply heart disease's existence. The pairwise method is first used to select variables and iteratively do the likelihood ratio test to find the best model. Using logistic regression with a logit link shows that seven factors significantly influence the odds of heart disease. These variables are sex (gender); cp (chest pain type); trestbps (resting blood pressure (in mm Hg on admission to the hospital); oldpeak (ST depression induced by exercise relative to rest); slope (the slope of the peak exercise ST segment); ca (number of major vessels (0-3) colored by fluoroscopy); and thal (whether the observation have detective). The largest difference occurs when comparing the odds of having heart disease between patients who have two major vessels colored by fluoroscopy and those who do not have major vessels colored by fluoroscopy. The patient with asymptomatic chest pain also has

huge differences in heart disease from the patient with typical angina. This indicates that the patient with two major vessels colored by fluoroscopy and asymptomatic chest pain might be the most severe factor of heart disease.

6.2 Limitations

There are several limitations of the model. Firstly, the dataset is too small. Logistic regression generally requires a large sample size, but there are only 297 samples for this research project. And in the case of relatively small data, the original data contains 14 variables. Lack of data problem may cause the model to overfit and affect the result. This can explain why all the main effects and interaction terms for the model selected by the stepwise method are insignificant. Because of the overfitting problem, the final model only contains the variables with significant main effects and no interaction terms are included.

Besides, only a small number of patients in Cleveland have been investigated, so the result from this project could not represent every social-demographic background. It is necessary to study more patients in multiple countries and get a more diverse sample for the relationship between different factors and heart disease.

6.3 Future Developments

In future developments, besides collecting more data from other countries, other predictor variables can be added such as do samples take long-term medications or whether they have regularity of work and rest. In addition, machine learning technique can be used to improve our dataset and results. A simpler classifier model can be used such as the Naive Bayes algorithm to analyse the data. To avoid overfitting problem, data with small size require a low complexity model. In this case, a simpler machine learning algorithm will be more suitable for small data sets (“5 Ways to Deal with the Lack of Data in Machine Learning,” n.d.). Besides, it is also possible to synthesise the data by using the Synthetic Minority Over-sampling Technique (SMOTE) or Modified-SMOTE. “SMOTE takes the minority class data points and creates new data points between any two nearest data points joined by a straight line.” (The Ultimate Guide to Synthetic Data, 2018). Synthetic data refers to data created using artificial techniques rather than collecting data from the real world.

Sometimes the actual data would have limitations so that Synthetic data can meet specific needs or conditions that actual data cannot achieve. It can also overcome the lack of real-world data problems and relieve the trouble of overfitting.

References

García-Portugués, E. (n.d.). 5.7 Model diagnostics | Notes for Predictive Modeling. Retrieved April 9, 2022, from https://bookdown.org/egarpor/PM-UC3M/glm-diagnostics.html

Logistic Regression Assumptions and Diagnostics in R - Articles—STHDA. (n.d.). Retrieved April 9, 2022, from http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

The Investopedia Team. (2022, February 8). *Variance inflation factor (VIF)*. Investopedia. Retrieved April 10, 2022, from https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factor%20measures%20how,standard%20error%20in%20the%20regression.

The Ultimate Guide to Synthetic Data: Uses, Benefits & Tools. (2018, July 19). https://research.aimultiple.com/synthetic-data/

Triveri, J. D. (2017, June 6). Goodness of Fit and Significance Testing for Logistic. Regression Models. The Pleasure of Finding Things Out. https://jtrive84.github.io/goodness-of-fit-and-significance-testing-for-logistic-regression-models.html

World Health Organization. (2022). *Cardiovascular diseases*. World Health Organization. Retrieved April 10, 2022, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

4.5 Multicollinearity | Practical Econometrics and Data Science. (n.d.). Retrieved. April 9, 2022, from http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html

5 Ways to Deal with the Lack of Data in Machine Learning. (n.d.). KDnuggets. Retrieved April 9, 2022, from https://www.kdnuggets.com/5-ways-to-deal-with-the-lack-of-data-in-machine-learning.html/