Assignment 2: Predicting Heart Attacks

Michael Beauchamp

UMGC Data 630: Machine Learning

Fall 2021

Dr. Firdu Bati

**Introduction**

Heart disease is a leading cause of death for both men and women in the United States. According to data provided by the CDC, in America a person dies from heart disease every 36 seconds and account for 655,000 deaths every year (Centers for Disease Control and Prevention, 2020). Dr. Robert J. Goldberg of the department of Cardiology at the University of Massachusetts Medical School conducted a study of cardiology patients to determine if certain variables contribute to their survival rates. The dataset and description of the variables were made available by Dr. Bati from the University of Maryland Global Campus on their internal site.

A linear regression model assumes that “there is approximately a linear relationship” between the dependent and independent variables, or as the dependent variable X moves along the line of best fit the Y value moves concurrently (James, Witten, Hastie & Tibshirani, 2021). On the other hand, a logistic model identifies the probability of the outcome falling into one of two categories. Similarly, a Naïve Bayes model can use a categorical dependent variable to determine probability. Utilizing the dataset provided by the study, I am going to create two models to demonstrate the statistical significance of the independent variables and the strength of the model in its predictive capabilities on the dependent variable. For this dataset, the dependent variable is binary, that is, the value in each record is either a 0 for alive or a 1 for deceased. Consequently, I will create a logistic regression and a Bayesian classification model using the same variables and compare their predictive capabilities on patient survivability.

**Analysis**

The first step of the analysis is to explore the dataset to gain an understanding of what the dataset provides in the way of independent and dependent variables. Using R studio, I loaded the heart attack study into a variable called *heart* using the *read.csv* command. A high-level view of the data is provided using the command *str(heart)* which gives the name of the variables, their datatypes and examples of the data which can be seen in figure 1 of the appendix. I immediately notice that there are abbreviations in the variable names that don’t make sense without context. The reference provided by Dr. Bati, appendix figure 2, explains the variables names in more detail. Some of the variables provided are the age of the patient at hospital admission, gender, vital statistics including heart rate, blood pressure and body mass index, history of heart attack, length of hospital stay and vital status at the last follow up. There are three date variables saved as factors with only the body mass index variable saved as a number. The other 18 variables are integer datatypes.

Running the *summary(heart)* command, as seen in figure 3 of the appendix, provides general statistical knowledge of the variables and if there are null values. Examining some of the variables, such as age, I see that the youngest patient in the study is 30, the oldest is 104 and the average patient is almost 70 years old. Most average body mass index is 26.61 while the minimum is 13.05 and maximum is 44.84. The *summary(heart)* command reveals that there are no null values in the data, however, because many of the columns are categorical with records having either a 0 or 1 value it leaves the *summary(heart)* command somewhat less useful as an exploratory tool.

Normally the next step is to visualize the data but due to their not being any null values and the way plotting works with *ggplot*, I decided to do data pre-processing now. To use some of the *ggplot* functionality I needed to have some of the binary variables changed into factors. I used the code *heart$gender<-factor(heart$gender)* and *heart$fstat<-factor(heart$fstat)* to convert the gender and fstat variables. I then deleted the columns id, admitdate, disdate and fdate from the data because they would be useful for time series but would create rules that I fell would be removed by the minimal adequate model. Finally, I binned the age variable by using the discretize command.

In figure 4 I plotted the gender of the patients to gain an understanding of the mix between male and female. In the study we have a 60-40 split favoring male patients. Exploring the gender variable even further, we can plot the survivability of the genders by including the variable fstat, which tells us if the patient is alive or deceased at the time of their last follow-up. In figure 5, we see that male survivability is good as the number of males alive at their last follow-up was significantly higher than deceased. However, female patients’ survivability remained relatively flat, with a smaller portion of the patients being decease at the time of their final follow-up. Figure 6 shows the spread of male and female patients by age group. Most of the male patients are in the age group 67-79.3 years old while many of the female patients are in the 79.3-91.7 age group.

Figure 7 is a comparison of the body mass index or BMI of the patients by their age and gender. The red line in the graph represents a BMI of 25%, which according to the CDC is the lowest BMI value in the overweight category (Centers for Disease Control and Prevention, 2021). For the graph, every patient above the red line is considered either overweight or worse, obese. The visualization of BMI shows most of the patients are above the healthy level as stated by the CDC, which logically means BMI is going to be an important variable in the predictive capabilities of the model. Another interesting aspect of the data that more than likely will affect the model is a patient’s history of cardiovascular disease. In figure 8, I plotted the history of cardiovascular disease of male and female patients in the study. Most patients in the study have a history of cardiovascular disease and this fact is true for both male and females.

The variable mitype indicates if the patient had a history of q wave in their electrocardiogram. According to a study by Rex N MacAlpin MD, “AQW were a strong indicator of organic heart disease in both adult age groups, but their utility to indicate MI was age-dependent” (MacAlpin, 2006). The study indicates the presence of abnormal q waves (AQW) could predict myocardial infarction (heart attack) but only for older patients. In figure 9, I plotted the age of the patients against the variable mitype and found that most of the patients in the study did not have a history of abnormal q waves. Contrary to Dr. MacAplin’s study, the only age group where more patients showed a history of AQW were in the younger patients. Checking this variable’s significance in the model will be interesting.

**Results**

**Logistic Model**

To start creating the logistic model, I refreshed the heart attack data by reading the csv file back into the heart variable. Then, I began to clean the data of unwanted columns such as the id column and all the columns with dates. These columns would be problematic in the model and don’t contribute to its predictive capabilities. The next step was to split the data into one group called *train.*data that houses 70% of the dataset and a variable *test.data* that contains the remaining 30% of the dataset. Running the code *model<-glm(fstat~.,family=binomial,data=train.data)* creates the model using fstat as the dependent variable. The model is trained on the 70% of the dataset that we stored into the *train.data* variable. Once the model is built, we can call the model descriptions using the *summary(model)* command as seen in figure 10. According to the results of the model age, year and lenfol are the most significant variables of the dataset on the dependent variable fstat because their p values are lower than .05.

A confusion matrix is a useful tool to determine the accuracy of the model. In figure 11, the confusion matrix tells us that the number of correctly classified instances is 328, 24 misclassified instances with 352 total instances used to create the model. To assess the accuracy of the model, we divide 328 by 352 and get a accuracy of 93%. Now that we know the accuracy of the model run on the test data, we can evaluate the model on the test data to see if the resulting confusion matrix shows similar accuracy. Using the code *mypredictions<-round(predict(model, test.data,type=”response”))* we store the predictions of the model into the variable mypredictions. Next, we run the same command for the confusion matrix, *table(mypredictions, test.data$fstat)*, and get the result in figure 12. From the confusion matrix on the test data, the number of correctly classified instances is 134, 14 misclassified instances with 148 total instances. The accuracy of the model evaluated on the test data is 134 divided by 148, or 90.5% accurate.

**Minimal Adequate Model**

The minimal adequate model is run to remove the variables that are not significant to the model one step at a time to maximize its simplicity while maintaining its accuracy. The model’s Akaike Information Criterion (AIC) is “a number score that can be used to determine which of multiple models is most likely to be the best model for a given dataset” (Zajic, 2019). Therefore, removing variables removes their corresponding AIC score from the total. The goal is to get the lowest AIC score possible. The original model contained 17 variables with a AIC score of 191.26. The minimal adequate model ended with just 5 variables and a score of 173.52, a significant reduction (see figure 13). To test the minimal adequate model, I re-ran it using only the variables that were left.

The summary of the new model shows that the intercept moved from 16.04 to 18.82 and the 3 most significant value of age, year and lenfol all had a decrease in their p values. In figure 14, I show the summary but also completed the confusion matrix to test the new model’s accuracy. In this case there were 331 instances classified correctly, 21 wrong and 331 total instances. If we divide 331 by 352 we get the accuracy of the model being 94% which is a percent higher than the initial model. Running the new model on the test data, we get a confusion matrix that says there were 139 successful instances, 9 wrong and a total of 148. If we divide 139 by 148, we get a accuracy value of 93.9% which is 4% higher than the first model tested on the test data.

**Naïve Bayes**

In order to run the Naïve Bayes model, the variables must be a factor datatype. To create the model the first step is to read the dataset back into the heart variable. Then, I completed the necessary steps to remove the id and date variables and converted the remaining variables into factors. Once that step was complete, it was time to split the data into groups again, with 70% of the data into the train variable and 30% into the test variable. Finally, I ran the code to create the model based off the train data and printed the results.

In figure 16 you can see the first variable results. The print command shows us that the probability of a patient being alive is 55.7% while the probability of the patient being deceased at their final check in is 44.3%. The first variable listed is age with a bin at 30-42.3 years old. If a patient is alive at their final check, that is fstat=0, then the probability that the patient is in the age range 30-42.3 is 4.1%. If the person is decease, of fstat=1, the probability that the patient is in the age range 30-42.3 is only .64%.

The confusion matrix for the training data, figure 17, tells us that the number of successful instances is 298, 54 were unsuccessful instances with a total of 352 total instances. Dividing the successful instances by the total, we get a accuracy of 84.6%. Also in figure 17, we see the accuracy of the model on the test data. The confusion matrix says that were 133 successful instances, 15 were unsuccessful with a total of 148 instances. Dividing the successful instances by the total, we get a accuracy of 89.8%. One final visualization that clearly shows the relationships between the predicted and actual values of fstat is the mosaic plot. The blue squares represent the portion of the instances where the predicted and actuals are the same while the red represents the misclassified instances. By looking at the chart in figure 18, you can quickly and clearly see that the model was highly accurate, otherwise there would have been significantly larger red squares.

**Conclusion**

In this exercise I created 2 models aimed at predicting the likelihood that a patient was alive or deceased at the time of their final check in. The initial logistic model that was created resulted in an accuracy of 93% including all the tested independent variables. However, we were able to simplify the model significantly by running the minimal adequate model, reducing the number of independent variables from 17 to 5 and increasing the accuracy too 94%. The Navie Bayes model that was created based on the same dataset proved to be inferior with an accuracy of only 89.9%. Hospital systems rely on data more and more every day. As our understanding of complex disease and viruses continue to evolve, so does our ability to recognize the importance of data collection. Not only does how much data matter, but which data is just as important. In an article written by Nicole Foster, Dr. Stephen Weng of the University of Nottingham and his team can “predict heart attacks 10 years before they occur with 76 percent accuracy” (Foster, 2018). Data sharing in health systems are also providing insight into patient care. For example, a retinal photograph can reveal issues with the blood vessels in the eye indicating diabetes, a common ailment of patients who suffer heart attacks.

Due to the age of the dataset, I would only take the results as a starting point for predicting heart attacks. The variables that we have learned contribute to heart disease could add a lot to this study. A similar dataset with those variables included could result in a more accurate model that could be used to save patient lives. In an article on the American Academy of Family Physicians website, Doctor Kyle Jones points out that a study found most of the preventative care methods don’t produce a reduction in operating costs, however, they are effective at saving lives (Jones, 2020). While hospital systems operate on slim margins, being able to predict catastrophic health issues years before they arise will help doctors and patients plan to mitigate the risk.

The dependent variable of this dataset lends itself well to a logistic model. Having a yes/no category means that a linear model would not be effect as it is better suited at predicting a continuous variable. The Naïve Bayes model, while also being able to predict variables is yes/no records, in this case did not provide a model as accurate as the logistic model. In future studies, I would recommend that health care experts include new variables into the data to test their significance on patient survivability while sampling a larger group of patients. Some contributing factors could also be missed if the study was done in just one location in the country. With the ability of our health systems to share data, this project could include patients worldwide. With the small sample we were able to create a model with accuracy of 94%, it would be interesting to see how the model would be affected by populations from all over the world. Once the model is complete, healthcare professionals could use that model and its variables as the starting point to discover many new says to diagnose and treat heart disease before it becomes too late for a patient to act.

**References**

Centers for Disease Control and Prevention. (2021, August 27). *About adult BMI*. Centers for Disease Control and Prevention. Retrieved September 18, 2021, from <https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html>.

Centers for Disease Control and Prevention. (2020, September 8). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved September 16, 2021, from https://www.cdc.gov/heartdisease/facts.html.

James, G., Witten, D., Hastie, T., Tibshirani, R. (August 4, 2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.)[PDF file]. 61. <https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf>.

Jones, K. (2016, April 12). *Does prevention save money? that's the wrong question*. AAFP Home. Retrieved September 20, 2021, from https://www.aafp.org/news/blogs/freshperspectives/entry/does\_prevention\_save\_money\_that.html.

MacAlpin, R. N. (2006, July). *Significance of abnormal Q waves in the electrocardiograms of adults less than 40 years old*. Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc. Retrieved September 18, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6932435/>.

Foster, N. (2018, July 23). *The future of heart attack prediction*. Mended Hearts. Retrieved September 20, 2021, from https://mendedhearts.org/story/the-future-of-heart-attack-prediction/.

Zajic, A. (2019, December 27). *Introduction to aic - akaike information criterion*. Medium. Retrieved September 19, 2021, from https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced.

**Appendix**

Text

Description automatically generated

Figure 1: The str(heart) command, exploring the datatypes of each of the 22 variables

Graphical user interface, text

Description automatically generated

Figure 2: Description of the variables provided by Dr. Bati

A screenshot of a computer

Description automatically generated with low confidence

Figure 3: The summary(heart) command run to get a general idea of the statistical properties of the variables

Chart

Description automatically generated

Figure 4: Distribution of male vs. female patients

Chart, bar chart

Description automatically generated

Figure 5: Survivability of male vs. female patients

Chart, bar chart

Description automatically generated

Figure 6: Ages of male and female patients

Chart

Description automatically generated

Figure 7: BMI of Male and Female patients grouped by age

Chart, bar chart

Description automatically generated

Figure 8: History of cardiovascular disease in males and female patients

Chart

Description automatically generated

Figure 9: History of Q Waves by patient age groups

Table

Description automatically generated

Figure 10: Summary of the logistic model created with all variables

A picture containing graphical user interface

Description automatically generated

Figure 11: Confusion matrix of the logistic model on the training data

Text

Description automatically generated with low confidence

Figure 12: Confusion matrix of the test data

A picture containing text

Description automatically generated

Figure 13: The Minimal Adequate Model

Text

Description automatically generated

Figure 14: Running the model with variables from the MAM, confusion matrix created

Table

Description automatically generated

Figure 15: Confusion matrix on the test data of the MAM

Text

Description automatically generated

Figure 16: code for splitting the data and creating the Naïve Bayes model, then printing the results

Graphical user interface, text

Description automatically generated

Figure 17: Confusion matrices for the Naïve Bayes model on the training and test data

Chart, treemap chart

Description automatically generated

Figure 18: Mosaic plot of the predicted vs actual fstat variable