621-FinalProject

Ahmed Elsaeyed

2022-11-30

Probability of Cardiovascular Disease

This research centers around building a model that predicts the probability of cardiovascular disease given some prior health information.

Abstract- Ahmed

Keywords - Ahmed

Introduction - Ahmed

Literature review - Krutika

Methodology - Alec

HERE, PUT IN EACH OF THE VARIABLES AND THEIR DEFINITIONS (MORE FLUFF FOR THAT PAGE COUNT!). SEE THIS LINK: https://rpubs.com/amitkapoor/data621finalprj

EXAMPLE TABLE YOU CAN USE TO PUT IN VARIABLES AND THEIR DESCIPTIONS

Experimentation and Results

Data Exploration

Our data set had 4,240 observations and 16 variables. `Sex`, `education`, `currentSmoker`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes` and `TenYearCHD` in the analysis carried out in this report are categorical variables, while the rest are numeric. The histograms for the numeric variables revealed that `BMI`, `glucose`, and `totChol` were skewed, and a Box-Cox transformation was undertaken for these variables in order to create new variables that contained the transformed versions of `BMI`, `glucose`, and `totChol`. An analysis into the correlation was done in order to remove variables that were highly correlated with another. Also, the data had several observations with missing data. The methodology that was undertaken to account for these missing values is shown in the "*Missing Values*" section. For comparison purposes, two different datasets were created; one containing the original data with observations with missing values omitted, and another containing manipulated data, where the data went through several transformations as explained in this report.

Multicollinearity Analysis

Correlation refers to the strength of a relationship between two or more variables. Problems arise when there significant correlation between two or more predictor variables within a dataset, as this can lead to solutions that are wildly varying and possibly numerically unstable and there is redundancy between predictor variables.

One solution to this is to drop one variable that is highly correlated with the other. Calkins explains that we can describe the correlation values between variables in the follwing manner:

- Correlation coefficients between 0.9 and 1 = very highly correlated
- Correlation coefficients between 0.7 and 0.9 = highly correlated
- Correlation coefficients between 0.5 and 0.7 = moderately correlated
- Correlation coefficients under 0.5 = low correlation

The correlations between the variables in the dataset were calculated. What was found was that the vast majority of the variables had a correlation that was less than 0.25. `sysBP` and `diaBP` had a correlation of 0.79, meaning that these two variables were highly correlated. One of these variables was removed from the dataset. When a binary logistic regression model with the original data was constructed, it was found that `sysBP` had a p-value of 0.00015 indicating a high degree of statistical significance, while `diaBP` had a p-value of 0.55. Therefore, the `diaBP` variable was removed from the dataset. The second highest correlation value after this was 0.38, and that was for the `BMI` and `diaBP` variables, but with the removal of the `diaBP` variable, and the fact that 0.38 would be considered a low correlation, `BMI` was kept in.

Missing Values

The dataset itself contains several observations containing missing values. The percentage of missing data for each of the predictor variables is shown in Table 1.

Table 2: Percentage of missing data for each of the variables that had missing data.

Ultimately, it was decided to remove the observations with the missing values instead of imputing. The reason for this is explained in the "*Removing Outlier Values*" section.

Data Preparation

Removing Outlier Values

The boxplots, summaries, and histograms that were generated revealed that some of these heart rates, diastolic, and systolic values as well as total cholesterol do not seem realistic. Imputing the values would not be appropriate because this is medical data and would harm the authenticity of the dataset, and so we have decided to simply remove the bad data.

I removed the diaBP variable, because it had too high a p-value in the first model and it was highly colinear with sysBP

- Peter _____

-Krutika -heartRate: we can probably remove the records over 100 because that is probably indicative of a stroke in-progress -diaBP/sysBP: we can remove the extreme measurements here as well for similar reasons max sys/dia = 160/100 min sys/dia = 100/65 -tolChol: we can probably remove the records over 500

Box-Cox Transformation for Skewed Variables

For the variables that exhibited skewness as explained in the "*Data Exploration*" section of this paper, A Modern Approach to Regression with R explains the following:

> . . . if the skewed predictor can be transformed to have a normal distribution conditional on Y, then just the transformed version of X should be included in the logistic regression model. (Sheather, 284)

In order to address this skewness and attempt to normalize these features for future modeling, we employed Box-Cox transformations. Because some of these values include 0, we replaced any zero values with infinitesimally small, non-zero values.

The λ's that were used to transform the skewed variables are shown on Table 3.

Table 3: λ's for transforming skewed variables to normal distribution.

Split Data Into Testing and Training

The data was into testing and training subsets such that 70% of it will be used to train, and 30% to test. The first row shows the split for the testing data while the second row shows the split for the training data. Tables 4 and 5 show the distributions of the testing and training data using the original and modified datasets.

Table 4: Distribution of testing and training data using original dataset

Table 5: Distribution of testing and training data using modified/transformed dataset

Building Models

Binary Logistic Regression Model with Original Data

A binary logistic regression model was generated using the original dataset. The observations with missing values were omitted. The final binary logistic regression takes on the following form:

$$\begin{aligned} logit(p) = -9.155 + (0.504 * Sex_{male}) + (0.066 * age) + \\ (-0.290 * education_2) + (-0.223 * education_3) + (-0.120 * education_4) + \\ (0.086 * currentSmoker_{Yes}) + (0.019 * cigsPerDay) + (0.123 * BPMeds_1) + \\ (0.982 * prevalentStroke_1) + (0.091 * prevalentHyp_1) + (-0.435 * diabetes_{Yes}) + \\ (0.002 * totChol) + (0.018 * sysBP) + (-0.005 * diaBP) + (0.02 * BMI) + \\ (-0.005 * heartRate) + (0.01 * glucose) \end{aligned}$$

Equation 1: Binary logistic regression model using original data.

In this model, the variables that had a p-value greater than 0.05 were `Sex age`, `cigsPerDay`, `sysBP` and `glucose`. The sex of the person definitely seems to play a significant factor in determining whether or not a person will develop heart disease. Figure XX revealed that men that were surveyed on average smoked more cigarettes per day than women. In fact, Figure XX reveals that within the sample, women on average do not smoke. Since smoking increases the formation of plaque in the blood vessels which leads to coronary heart disease according to the CDC, the fact that men smoke more than women is why the sex variable in particular is of great statistical significance. The National Institute on Aging points out the following:

> People age 65 and older are much more likely than younger people to suffer a heart attack, to have a stroke, or to develop coronary heart disease (commonly called heart disease) and heart failure. Heart disease is also a major cause of disability, limiting the activity and eroding the quality of life of millions of older people. (NIH)

This explains why `age` was statstically significant in this model as well. The Brisighella Heart Study conducted in 2003 was able to conclude that within their sample that systolic blood pressure was a strong predictor of coronary heart disease, while diastolic blood pressure was not statistically significant. This result in this study is also seen in this model; `sysBP` has a much lower p-value (0.0001) than `diaBP` (0.55). `cigsPerDay` is tied to `Sex` as shown on Figure xx, which is why the `cigsPerDay` variable is also statistically significant. Park was able to conclude the following about glucose levels and coronary heart disease based on a study which had a sample size of 1,197,384 adults:

> Both low glucose level and impaired fasting glucose should be considered as predictors of risk for stroke and coronary heart disease. The fasting glucose level associated with the lowest cardiovascular risk may be in a narrow range. (Park)

This provides evidence as to probably why `glucose` is also statistically significant within the model.

Binary Logistic Regression Model with Modified Data

A binary logistic regression model was generated using the modified dataset. The observations with missing values were omitted. The final binary logistic regression model that used modified data takes on the following form:

$$\begin{aligned} logit(p) = -128.3 + (0.52 * Sex_{male}) + (0.073 * age) + \\ (-0.105 * education_2) + (-0.55 * education_3) + (-0.074 * education_4) + \\ (0.277 * currentSmoker_{Yes}) + (0.013 * cigsPerDay) + (0.091 * BPMeds_1) + \\ (0.651 * prevalentStroke_1) + (0.145 * prevalentHyp_1) + (-0.026 * diabetes_{Yes}) + \\ (0.009 * totChol) + (0.013 * sysBP) + (0.18 * BMI) + (-0.009 * heartRate) + \\ (0.013 * glucose) + (-13 * tf\_BMI) + (-140.9 * tf\_glucose) + (-1.39 * tf\_totChol) \end{aligned}$$

Equation 2: Binary logistic regression model using modified data.

The parameters in this model that had p-values less than 0.05 were `Sex`, `age`, `education_3`, `sysBP` and `glucose`. None of the Box-Cox transformed variables had p-values less than 0.05. We see the same variables (except `cigsPerDay`) that were statistically significant when the original data was used.

Step-AIC Binary Linear Regression Model with Original Data

Step AIC works by deselecting features that negatively affect the AIC, which is a criterion similar to the R-squared. It selects the model with not only the best AIC score but also a model with less predictors than the full model, since the full model may have predictors that do not contribute or negatively contribute to the model's performance. The direction for the Step AIC algorithm was set to `both`, because this implements both forward and backward elimination in order to decide if a predictor negatively affects the model's performance. The final step-AIC binary logistic regression model that used the original data takes on the following form:

$$\begin{aligned} logit(p) = -9.589 + (0.503 * Sex_{male}) + (0.072 * age) + (0.02 * cigsPerDay) + \\ (1.004 * prevalentStroke_1) + (0.017 * sysBP) + (0.023 * BMI) + (0.009 * glucose) \end{aligned}$$

Equation 3: Step-AIC binary logistic regression model using the original data.

The p-values for this model, which is far more parsimonious than the original binary logistic regression model that used all of the predictors, were all below 0.05 except for `prevalentStroke1` and `BMI`, which had p-values of 0.0861 and 0.1179 respectively.

Step-AIC Binary Logistic Regression Model with Modified Data

The final step-AIC binary logistic regression model that used the modified data takes on the following form:

$$\begin{aligned} logit(p) = -14.41 + (0.52 * Sex_{male}) + (0.073 * age) + (-0.09 * education_2) + (-0.55 * education_3) + \\ (-0.056 * education_4) + (0.02 * cigsPerDay) + (0.015 * sysBP) + (0.2 * BMI) + \\ (0.009 * glucose) + (-14.76 * tf\_BMI) \end{aligned}$$

Equation 4: Step-AIC binary logistic regression model using the modified data.

The p-values for this model, which is far more parsimonious than the original binary logistic regression model that used all of the predictors, were all below 0.05 except for `Intercept`, `education2`, `education4`, `BMI` and `tf_BMI`; 3 more than the amount of predictors that had values above 0.05 for the step-AIC binary logistic regression model using the original predictors.

Model Selection

The various metrics that were used to determine the best model are shown in Table 6. Figure XX is a bar chart which plots all of the metrics for all of the models. Table 6 and Figure xx reveal that the accuracy, precision, recall, AUC, and F-score are higher when the original data was used as opposed to the modified. Table 6 revealed that the step-AIC binary logistic regression model using the original data has the second highest AIC, but it also has the highest precision and F-score out of all of the models. Taking all of this into consideration, and the fact that the step-AIC binary logistic regression model using the original data is the most parsimonious out of all of the models, this model is the best performing model out of all of them. The AUC curve for the step-AIC binary logistic regression with original data is provided in Figure xx.

Table 6: Model metrics for binary logistic regression models

Discussion and Conclusions

References

Multicollinearity - Wikipedia: https://en.wikipedia.org/wiki/Multicollinearity#Consequences_of_multicollinearity

Correlation Coefficients - Keith G- Calkins: https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm

A Modern Approach to Regression with R - Simon Sheather

Smoking and Cardiovascular Disease - CDC: https://www.cdc.gov/tobacco/sgr/50th-anniversary/pdfs/fs_smoking_CVD_508.pdf

Heart Health and Aging - NIH: https://www.nia.nih.gov/health/heart-health-and-aging

The relationship between systolic blood pressure and cardiovascular risk–results of the Brisighella Heart Study - National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/12556653/

Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases - Chanshin Park: https://pubmed.ncbi.nlm.nih.gov/23404299/

Appendix

FIGURES AND CODE GO HERE.