Stats 101C Final Report

Prediction of Existence of Heart Disease

Bailey Kui, Christy Hui, Shirley Tang, Chae Yeon Lim
Lecture 2

Abstract

The main objective of this project was to use various supervised machine learning techniques to determine if a patient is diagnosed with heart disease. In this report, we will explore the given data set, explain the process of organizing the data, describe the results of the used models, give possible explanations on why the models acted the way they did, and provide suggestions on how to possibly yield a more consistent and stronger result.

For a more brief explanation on the process, this project had four main steps: explore and interpret the data, clean and modify damaging data points, generate models using the improved data set, and compare the model accuracy rates.

It was found that various observations consisted of missing or confusing data observations. Problematic numerical variables were imputed with the mean of other same-column observations while categorical variables were imputed with the mode of other observations. After trying Linear and Quadratic Discriminant Analysis, K-Nearest Neighbors, Random Foresting, and Logistic Regression models, we found that the latter yielded the most accurate results, with a public testing accuracy rate of 81.264% and total accuracy rate of 79.742%.

I. Introduction

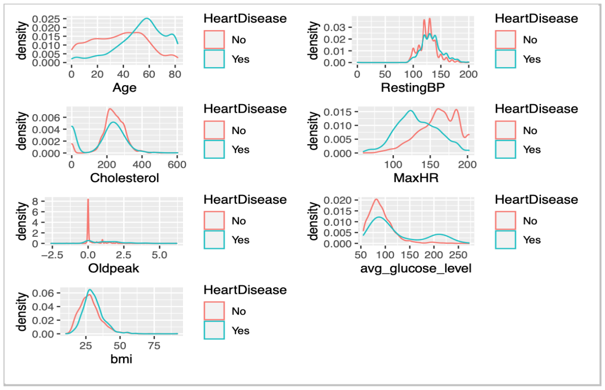
Heart Disease, an umbrella term for a multitude of heart conditions such as Coronary Artery Disease (CAD), Heart Failure, and Heart Arrhythmias, is the leading cause of death in the United States. The disease is even more dangerous for women, as it is responsible for 1 in every 4 female deaths. In a study conducted by "The Doctors Company," a medical malpractice insurer, patients die when their heart conditions are not correctly diagnosed in 70% of claims. Diagnosing heart disease is difficult, even for doctors, as symptoms can vary between women and men; furthermore, almost two-thirds of women who die suddenly from heart disease have no previous symptoms. This seemingly unpredictability of the disease can lead to difficulty diagnosing a patient correctly. Thus, finding an efficient and accurate model that can predict heart disease in a patient more accurately than a primary care physician or other doctors is paramount in saving millions of lives.

The provided dataset was split into two separate files: training and testing data. The training dataset contained 4220 observations and 20 columns representing explanatory information about the patient: for instance, the "Sex" variable was useful in determining the patient's biological sex, and "smoking_status" helped to discover if a patient smoked. Most importantly, the "HeartDisease" column contained information about whether or not a patient had heart disease ("Yes" or "No). The testing dataset contained 1808 observations and all of the same columns as the training data, excluding the "HeartDisease" response column. Since this project was conducted in the form of a competition, the "HeartDisease" column was absent in the testing data set. In order to find the best model that reliably predicts heart disease, each model's predictions were submitted into a third-party website that contained the testing data set's "HeartDisease" column.

II. Exploratory Data Analysis and Cleaning Data

Numerical

There were 7 numerical predictors total, including "Age," "RestingBP," "Cholesterol," "MaxHR," "Oldpeak," "avg_glucose_level," and "bmi". To visualize the distribution of these variables, we began by creating density plots of



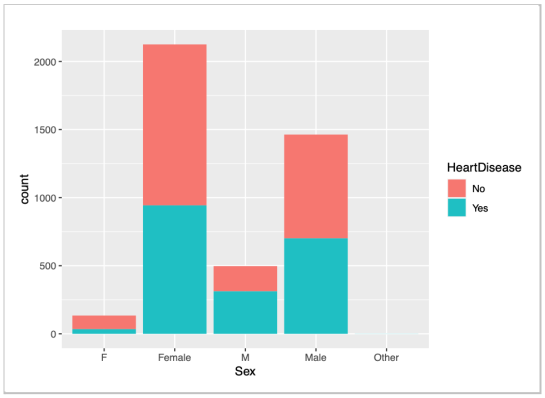
each numerical predictor, colored by the “HeartDisease” variable.

We saw that some predictors like “MaxHR” had differing distributions based on heart disease while others had less of a distinction, suggesting that certain variables would be more useful than others in classifying patients. In addition, we could see that “Oldpeak” had an unusually tall peak around 0.0 for those without heart disease, and “Cholesterol” had a similar indication of values at 0.

Categorical

There were 13 categorical variables in the dataset including the response variable: “Sex”, “ChestPainType”, “FastingBS”, “RestingECG”, “ExerciseAngina”, “ST_Slope”, “hypertension”, “ever_married”, “work_type”, “Residence_type”, “smoking_status”, “stroke”, and “HeartDisease.” For the categorical data, we utilized stacked bar charts of each variable based on “HeartDisease” to examine the levels of each variable, as well as to note any other significant patterns within the data.

One interesting thing we found in particular was that the “Sex” variable had overlapping levels (“F” and “Female”, “M” and “Male”) that could be recoded to two levels to help with the classification.



Cleaning

While inspecting the data observations, it became evident that the dataset was incomplete. In both training and testing data sets, multiple columns had missing values. These consistent offenders were the “ever_married,” “work_type,” and “Residence_type,” and “smoking_status” columns. The training data set had 2524 missing values and the testing data set had 1148 missing values

One solution to solve this problem of missing values is to impute them with the mode of existing values. Since the columns with missing values were all categorical, this process was fairly simple. For example, since the “ever_married” column contained more “Yes” answers than “No,” any missing values were replaced with a “Yes.”

The dataset was not only incomplete but also confusing. The numeric columns “Cholesterol” and “Oldpeak” included many zero values that were also seen from the density plots, which do not make sense in the context of the problem. To combat this nonsensical finding, we replaced the zero values with the mean of the existing values.

III. Methodologies

Choosing Predictors

While choosing the correct type of model is always important, it is also crucial to choose the correct variables to use when initializing said models. Another method is to perform logistic regression with all predictors to obtain the most significant predictors from that model. In this project, logistic regression was performed twice: once with uncleaned data and another with cleaned data. Some common predictors between the four methods described above include “Sex,” “ChestPainType,” “Cholesterol,” “FastingBS,” “MaxHR,” “Oldpeak,” and “stroke.”

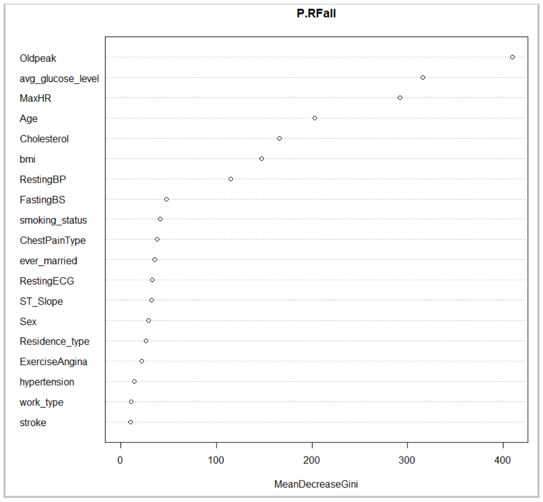
Linear Discriminant Analysis, Quadratic Discriminant Analysis, and K-Nearest Neighbors

The three worst performing models were Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (KNN) using all numeric predictors. Note that the numeric variables were standardized before performing these methods; however, the methods had difficulty in scoring a testing accuracy rate above 80% despite the fact that cross-validation was also performed on them. The best accuracy rates reported on Kaggle for LDA, QDA, and KNN were 0.79189, 0.79123, and 0.76126 respectively.

LDA and QDA may have performed inadequately because the assumptions of the models do not hold. Furthermore, a naive Bayes' classifier may be insufficient in predicting heart disease. Since KNN is a distance-based algorithm, it may have performed inadequately because the cost of calculating the distance between two points is too high. Furthermore, there are many numeric predictors (7). It is possible that the curse of dimensionality made KNN perform a lot worse than if we used a smaller amount of predictors. Finally, it may simply be that it is difficult to predict heart disease using the given predictors; thus, no matter what algorithm or model we use, the success rate of predicting heart disease is only around 80%.

Logistic Regression

For our best Logistic Regression model, we removed all the NA values and used the predictors: "Sex", "ChestPainType", "Cholesterol", "FastingBS", "MaxHR", "ExerciseAngina", "Oldpeak", "Avg_glucose_level", and "stroke" based on backward stepwise regression. We used a tuning parameter of .524, obtained by testing values using a for-loop. The accuracy reported on Kaggle was 81.26%. We also utilized the ROCR package to find the best combination of predictors with the best classification tuning parameter of .5134. The following model resulted in the same accuracy rate.



Random Forest

Random Forest is excellent with outliers and missing values, but ultimately we cleaned the data using mean or mode imputation. We utilized the variable importance plots to determine the best predictors to make the random forest model. We decided to use 4 predictors, because the square root of the total number of predictors (19) is generally a good number of predictors to use for any model.

IV. Discussion and Suggestions

A further look into literature and the performance of our model would not only give us insight on its application in a real life setting, but also inform us of other factors that may contribute to heart disease. Key questions that remained after our analysis include how our best model compares to the accuracy rate of doctors predicting heart disease in patients and whether doctors accurately predict the existence of heart disease more than 81% of the time. Recommendations for further exploration include analyzing the incorrect predictions, finding a commonality among all the observations, and taking note of what sets them apart from the observations that were predicted correctly. Another suggestion is to collect more data about each patient pertaining to ethnicity, ongoing family history of heart disease, or any other physiological data, and to experiment with better variable-transforming and modeling techniques to raise the accuracy of the model.

V. Conclusion

Overall, logistic regression was the best technique to predict whether a patient has heart disease or not based on an accuracy rate of 81.26% and using “Sex”, “ChestPainType”, “Cholesterol”, “FastingBS”, “MaxHR”, “ExerciseAngina”, “Oldpeak”, “Avg_glucose_level”, and “stroke” as predictors selected from backward stepwise regression.

Though we were able to look into changing the cut-off value for classifying patients as having heart disease or not using our logistic regression model, it would be interesting to see if we could apply any additional transformations to the predictor variables themselves, such as splitting numerical predictors into categories based on literature.

References

Almohalwas, Akram. *"Prediction of Existence of Heart Disease"*. October 26, 2021.

Heart Disease: Types, Causes, and Symptoms. WebMD. June 14, 2021, https://www.webmd.com/heart-disease/heart-disease-types-causes-symptoms.

Murphy, S. L., Xu, J., Kochanek, K. D., & Arias, E. *Mortality in the United States*. Centers for Disease Control and Prevention. November 29, 2018, https://www.cdc.gov/nchs/products/databriefs/db328.htm.

Taking the risks to heart: Misdiagnosis of heart disease. American College of Cardiology. February 20, 2017, https://www.acc.org/membership/join-us/benefits/additional-member-only-benefits/acc-and-the-doctors-company/the-doctors-company-updates/2017/02/20/12/55/taking-the-risks-to-heart-misdiagnosis-of-heart-disease.