MIS373 - Advanced Analytics Programming
Group 9: Evan Lock, Prerana Amargol, Anchit Nath, Willy Resendiz Torres, Robert Carstens

Description of project goals

Our group chose to analyze a data set about heart disease found on Kaggle. Heart disease refers to various conditions including heart attacks, arrhythmia, and heart failure in extreme cases. The dataset[1] used for this project is unique as it has a relatively large sample size and detailed information on various clinical and demographic variables. These variables include age, blood pressure, cholesterol, and lifestyle choices such as smoking and education. The dataset is sourced from research articles done by the American Heart Association (AHA), and Centers for Disease Control (CDC). The bulk of the data comes from voluntary surveys done through the National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS), which spans from 1999-2022. The cases were sourced from a wide array of patients as the surveys use a multi-stage probability design to select representative samples of the population. Given the dataset, the question we aimed to answer was which features given by the dataset have the greatest impact on whether the patient had a stroke.

Analyzing heart disease data to identify predictors is an important problem as heart disease remains a leading cause of death worldwide[2]. In addition, heart disease can decrease the quality of life, limit lifestyle choices, and exacerbate other conditions. The significance of this problem lies in its potential to provide insights into the risk factors associated with heart disease, which can aid prevention and treatment strategies. By identifying predictors of heart disease, this project could contribute to developing more accurate prediction models and personalized interventions, leading to improved patient outcomes and reduced healthcare costs. This work could be of interest not only to medical professionals but also to individuals who may be at risk of heart disease, as it could empower them to take proactive steps to reduce their risk. Ultimately, solving this problem could have a positive economic impact by reducing the burden of heart disease on healthcare systems and society as a whole.

---

[1] Heart Disease Dataset, Kaggle
https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset
[2] The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health
https://doi.org/10.1016/j.jacc.2022.11.005

Exploratory analysis

We first started by examining the data itself. The Kaggle dataset has 4238 records and 16 features. The features include basic characteristics including gender, age, and education, as well as more specific risk factors including hypertension, blood pressure, and smoking habits. The target variable is a feature called Heart stroke that is either *yes* or *no*. Before we began exploratory analysis, we had to do some pre-processing on the data. First, we dropped records with null values from the data frame. This left us with 3656 records. Next, we changed any variables that had values of yes and no into binary variables of 1 and 0. This included making a new target variable (called *target*) based on the Heart_stroke variable.

Our next point of analysis was to see what proportion of the data has heart disease. We did a value count on the target variable to find that of the 3656 records, 3099 didn't have heart disease. In other words, only 15% of this data set experienced heart disease. This was an important statistic for us to understand early on in our analysis so we could compare normalized versions of this imbalance data set.

One basic fact about this data is that the ages range from 30 to 70 years, with the majority of the records aging from 40-50 years old, with the mean age being around 49 years old. We also saw that as age increased, the proportion of the population experiencing heart disease increased as well. For this data, people under 40 had a 5% heart disease rate and people over 60 had a 30% heart disease rate. We also plotted heart disease by education level, which interestingly revealed that uneducated people had higher levels of heart disease. This could be tied to lack of education about mitigators of heart disease including nutrition and exercise.

Next, we compared the summary statistics of the numeric variables including age, cigarettes per day, cholesterol, and more for those with and without heart disease. Unsurprisingly, all numeric variables were higher for those with heart disease than without. The most significant differences were present in the age variable (48 years old vs 54 years old), cholesterol variable (235 units vs 246 units), and glucose (80 units vs 88 units).

Solution and insights

To start the regression task, we created a design matrix with all available features (specifying that Gender and Education were categorical variables) and that the rest were numerical. The target variable was a binary variable based on whether the patient had experienced a heart stroke. After fitting a logistic regression, we found that the variables with the largest

absolute value of coefficient weight were: prevalentHyp, C(Gender)[Female], C(education)[T.primaryschool], C(Gender)[Male], C(education)[T.uneducated]. Our testing accuracy on this model was 0.8605287 with a baseline accuracy of 0.8587055.

Intuitively, it makes sense that prevalent hypertension had a large magnitude in predicting heart disease. However, it was surprising how impactful features such as gender and education were. Our group was expecting a heavier impact from features such as blood pressure and glucose, which had very low feature weights for this model. One cause of this discrepancy between our prior understanding of heart disease and the results of the model may have been the scales of the data. For example, the coefficient of the cigsPerDay variable might have been lower since it's not a binary variable (like a lot of the other data). To address this, we decided to normalize our data, so all numeric values were on a scale of 0 to 1. In order to normalize the model, we changed all non-binary variables by subtracting the minimum value from the cell value and dividing the result by the range of the column.

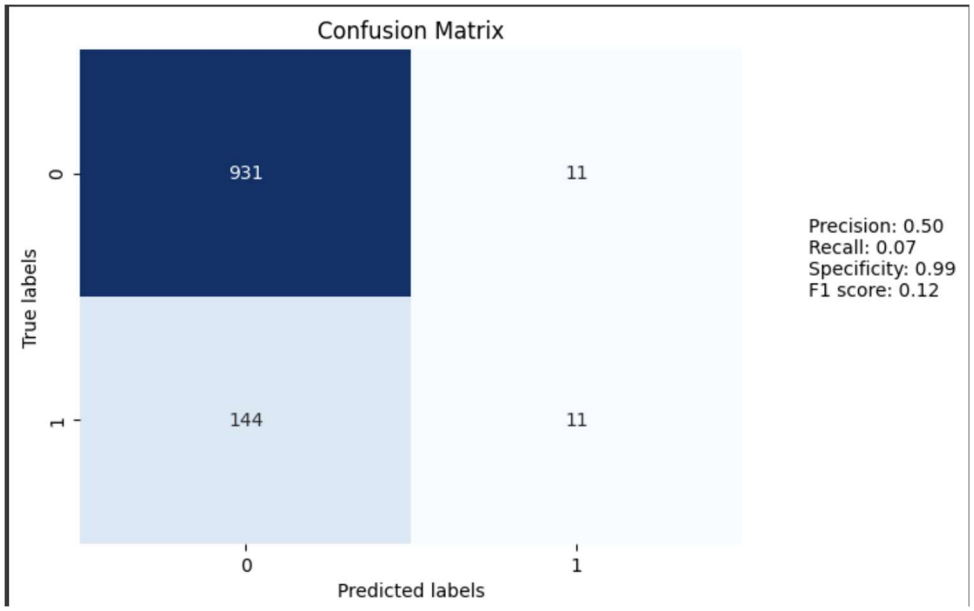
After re-running the model on the normalized data with all the features, we got a lower testing accuracy of 0.8587055, but a model with different feature weights. Our normalized model yielded the following as the most impactful features: age, sysBP, glucose, totChol, cigsPerDay. These features were much more consistent with our understanding of heart disease risk factors. The least impactful features were: C(education)[T.uneducated], C(education)[T.postgraduate], C(education)[T.uneducated], currentSmoker, prevalentHyp. It makes sense that education wasn't an impactful feature, although it was still surprising because, in our exploratory analysis, we seemed to see a correlation between education levels and proportion of people with heart disease.

After declaring these the most impactful variables, we re-ran the model with only these 5 features as part of the design matrix. With only these 5 features, our testing accuracy increased to 0.861440. Adding a 6th highest weighted variable decreased our testing accuracy again, so we decided to use the 5 features listed above in our model.

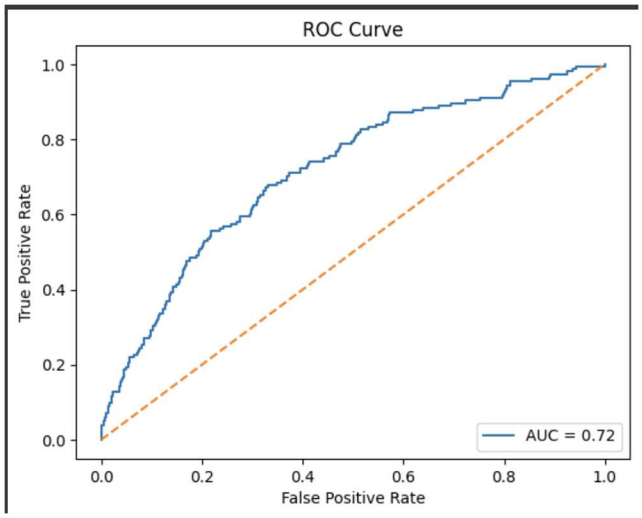
Based on the results of the normalized model with all features included, it appears that age and systolic blood pressure (sysBP) are the most important predictors of cardiovascular disease. Both of these features have odds ratios greater than 8. Other important features include glucose, total cholesterol (totChol), and smoking (cigsPerDay). These features have odds ratios between 2 and 5, indicating a moderate to strong association with cardiovascular disease.

Interestingly, some features that are commonly associated with cardiovascular disease, such as BMI and heart rate, have relatively low odds ratios in this model. This may be due to the fact that these features are highly correlated with other predictors, such as age and blood pressure, and may not add much additional information to the model.

The confusion matrix reveals that the model correctly identified 11 individuals with the target condition (true positives) and correctly identified 932 individuals without the target condition (true negatives), but it also incorrectly identified 11 individuals as not having the target condition when they did (false negatives) and incorrectly identified 144 individuals as having the target condition when they did not (false positives). The model's accuracy is 0.856, precision is 0.500, recall is 0.071, specificity is 0.988, and the F1 score is 0.125, indicating a relatively low overall performance. While the model classified a large proportion of cases correctly, a high number of false positives is not desirable, especially in medical diagnoses.



Additionally, in our logistic regression model, the AUC value is 0.72, which indicates that the model is performing better than random guessing. The ROC curve shows that the model can distinguish between positive and negative examples with some degree of accuracy, but there is a trade-off between sensitivity and specificity.



These findings have important implications for healthcare providers and individuals looking to reduce their risk of heart stroke. Healthcare providers can use this information to identify high-risk patients and provide targeted interventions to reduce their risk of developing heart stroke. Individuals can also use this information to make lifestyle changes that can help reduce their risk, such as quitting smoking, improving blood pressure control, and maintaining a healthy BMI. The data set's limitations include a relatively small size, lack of diversity, limited number of features, and possible missing or incomplete data. These factors may restrict the scope and generalizability of the analysis, as well as the model's predictive power. Moreover, the use of a logistic model may oversimplify the relationship between predictors and the outcome variable, which could affect the validity of the conclusions drawn from the analysis.