Team 2: Jacqueline Jia, Madeline Wood, Kitty Chen

CHL5230H Fall 2023: Datathon #2 Written Report

Introduction

Heart disease and stroke are two leading causes of mortality worldwide (Tsao et al., 2023). Up to 70% of adverse cardiovascular events could be prevented by targeting a small cluster of modifiable risk factors, including diabetes, obesity, smoking, and high blood pressure (Yusuf et al., 2020). Demographic factors such as socioeconomic status, sex, and age further influence an individual's risk for developing cardiovascular disease, with older individuals at particularly greater risk than their younger counterparts (Yusuf et al., 2020). In this study, we used machine learning to investigate whether a range of clinical, behavioral, and demographic factors can predict the risk of adverse cardiovascular events (i.e., heart disease or stroke) in adults over 40 years of age.

Data Engineering Process

Our team used the 'Cardiovascular Event Dataset'. Explanatory variables included gender, age, lifetime marital history (i.e., ever married vs. never married), work type, residence type (i.e., rural vs. urban), average glucose level, hypertension, body mass index (BMI), and smoking status. Our outcome of interest was the occurrence of heart disease and stroke, which we coded as a binary variable whereby history of either heart disease or stroke = 1 and history of neither = 0. Given that younger adults are substantially less likely to experience adverse cardiovascular events (Dhingra & Vasan, 2012), we restricted our analysis to middle-aged and older individuals (i.e., >40 years of age). There were missing values for BMI (N = 201) and smoking status (N = 1,544). Missing values for BMI were replaced with sex-specific sample median values. For smoking status, missing values were randomly assigned as one of the three available options (i.e., formerly smoked, smokes, or never smoked). We compared the distribution of the smoking data before and after replacing the missing values to ensure they were similar. We calculated descriptive statistics (e.g., means, proportions) and examined the distribution and missingness for each variable in the dataset. Categorical variables (i.e., occupation, residence type, gender, ever married, and smoking status) were converted into dummy variables using 'one-hot' encoding, whereby each categorical option was coded as a new binary variable where 1 = true and 0 = false. Continuous features were standardized using the means and standard deviations of the sample data. A correlation heat map was used to check for multicollinearity in features. Since we modeled the occurrence of heart disease or stroke as our outcome, we removed individual features for heart disease and stroke from the dataset before running the logistic regression model. We also removed the participant ID from the feature set.

Analysis

We used a logistic regression model with lasso regularization to predict the probability of experiencing an adverse cardiovascular outcome (i.e., heart disease or stroke) using the above-described features. We selected logistic regression because we were interested in classifying probability for a binary outcome. Lasso regularization was applied to the model for optimal feature selection and to prevent overfitting with noise in the training set. The model was

trained using an 80-20 train-test split. In the logistic regression model, class weights were balanced to account for outcome imbalance (467 participants with outcome vs. 2,399 without; 16.3%). A range of model evaluation techniques were used to evaluate model performance. Specifically, we used a confusion matrix to compare the actual and predicted classifications of the binary outcome. A classification report was used to summarize the performance in terms of precision and accuracy for each class. A model summary report measures how well the model fits the training data. The ROC curve evaluates how well it distinguishes the binary classes, in which the Area Under the Curve (AUC) ranges from 0 (poor classification) to 1 (correct classification).

Findings

We included *N* = 2,866 adults aged >40 years of age in our analytic dataset (mean (*SD*) age = 60.2 (12.1 years), *N* (%) female = 1,675 (58.4)). Our model had an overall accuracy of 68%. Based on the classification report, the model was better at correctly classifying individuals without heart disease or stroke (F1 = 0.78) than it was at correctly classifying individuals with heart disease or stroke (F1 = 0.41). In the confusion matrix, we observed that our model identified more true positives (*N* = 324) than true negatives (*N* = 64) and more false positives ((*N* = 158) than false negatives (*N* = 28). This was also confirmed with an AUC of 0.68, in which there is a higher chance of positive class values such as true positives and true negatives than negative class values. The pseudo R^2 for the logistic regression model was 0.14, suggesting that our model explains approximately 14% of the variance in the probability of heart disease or stroke.

Conclusion

To sum up, we developed a logistic regression model that used standard clinical and demographic factors to predict an individual's risk of experiencing heart disease or stroke with reasonable accuracy. From a clinical perspective, this model could help identify individuals who may be at elevated risk of adverse cardiovascular events. From a public health perspective, our findings could inform the development of interventions targeting these modifiable risk factors (e.g., BMI, hypertension, diabetes) to lower cardiovascular risk. Strengths of our study include the large sample size and the integration of machine learning techniques. There are also some limitations to note, including the low accuracy for identifying positive cases, which may be related to the imbalanced outcome distribution. Also, our low pseudo-R^2 suggests that there are many factors beyond those included that might influence individual risk for adverse cardiovascular events. Future research should leverage larger datasets with more diverse clinical, behavioural, environmental, and genetic features to develop models that can more accurately estimate the risk for heart disease and stroke.

Individual Contributions: JJ, MW and KC contributed to the conceptualization of the study, analysis and interpretation of results. JJ created the code for the model, MW and KC prepared the final report, and all contributed to the preparation of the presentation.

Code and Presentation

Team 2: Jacqueline Jia, Madeline Wood, Kitty Chen

References

Dhingra R, Vasan RS. Age as a risk factor. *Med Clin North Am*. 2012;96(1):87-91. doi:10.1016/j.mcna.2011.11.003

Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. *Circulation.* 2023;147(8):e93-e621. doi:10.1161/CIR.0000000000001123

Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet*. 2020;395(10226):795-808. doi:10.1016/S0140-6736(19)32008-2