Applications in Machine Learning to Predict Coronary Heart Disease

Zachary Katz (zak2132), Hun Lee (sl4836), and Tucker Morgan (tlm2152)

5/12/2022

Contents

Introduction

Heart disease accounts for roughly 695,000 fatalities annually in the United States alone, with known risk factors that include high cholesterol, smoking, and blood pressure. In this project, we aim to utilize observations from the Framingham cohort study to predict whether a particular study subject will or will not develop coronary heart disease in the next decade. Our data subset of longitudinal observations come from Kaggle; its cleaning generally entailed converting appropriate variables to factors (and re-leveling where needed), renaming and recoding binary 1/0 variables with more descriptive "yes" and "no" for ease of interpretation, and excluding observations with missing data on any measure. Prior to such exclusions, the full set of data contained 4,240 total observations across 16 variables, which are:

7 categorical predictors: `sex` (self-reported subject sex that takes values "male" or "female"); `education` (study participant's education that takes ordinal, mutually exclusive values "some_HS" (some high school completed), "HS_grad" (completed high school but did not attend college), "some_college" (attended college but did not graduate), and "college_grad" (graduated university)); `current_smoker` (binary "yes" or "no" indicating whether the participant was a smoker at the time of physical examination); `bp_meds` (binary "yes" or "no" indicating whether the participant was using anti-hypertensive medications at the time of physical examination); `prevalent_stroke` (binary "yes" or "no" indicating whether the participant had experienced stroke by the time of physical examination); `prevalent_hyp` (binary "yes" or "no" indicating whether the participant was being treated for active hypertension at the time of physical examination); and `diabetes` (binary "yes" or "no" indicating whether the participant was diagnosed as diabetic according to pre-specified criteria at the time of physical examination).

8 continuous numeric predictors: `age` (age in years at the time of medical examination); `cigs_per_day` (average number of cigarettes smoked each day at the time of medical examination, notably not conditioned on smoking status (i.e. for those with `current_smoker` status as "no", should take the value 0)); `tot_chol` (total blood cholesterol in mg/dL at the time of physical examination); `sys_BP` (systolic blood pressure in mm Hg at the time of physical examination); `dia_BP` (diastolic blood pressure in mm Hg at the time of physical examination); `bmi` (body mass index in kg/m^2 in mm Hg at the time of physical examination); `heart_rate` (resting heart rate in beats per minute at the time of physical examination); and `glucose` (blood glucose level in mg/dL at the time of physical examination).

Finally, beyond our 15 covariates lies our outcome (response) variable `ten_year_chd`, which is a binary indicator for the presence or absence of coronary heart disease (CHD) at 10 years of follow-up. Of our 4,240 observations, 3,596 (84.8%) have absence of CHD, whereas 644 (15.2%) have CHD present – a notable class imbalance.

EDA Figures

Notably, 582 rows (13.7% of our observations) lacked at least one data point, with `glucose` accounting for the plurality of missing data (9.15% missing rate). Moving forward, we assume that our data is missing at random, and consequently build a KNN imputation step (using five nearest neighbors) into our preprocessing functionality, which also includes centering, scaling, and BoxCox transformations where possible.

When we stratify the distributions of our continuous predictors by outcome status, we find the most substantial differences in median age and systolic blood pressure for those with and without CHD at 10 years. Looking at the proportions that have CHD present or absent across levels of our factor variables, there appears to be the most substantial differences for stroke history, diabetes status, and blood pressure medication status.

We observe no major multicollinearities, with the highest correlations (all sub-0.80) found between systolic and diastolic blood pressure, cigarette smoking and cigarettes smoked per day, prevalent hypertension and blood pressure, and glucose levels and diabetes comorbidity. In addition, CHD status is most correlated with age, systolic blood pressure, and prevalence of hypertension, which is unsurprising given our prior exploratory visualizations stratified by CHD class.

Modeling

In prior work, we attempted to rectify class imbalance through upsampling and downsampling. However, because this method was ineffective in improving our AUC, the focus here was on trying additional models, including ones with increased flexibility. We began by splitting the data into 80% training set and 20% test set. Given lack of major collinearity and KNN imputation for missing data points, all available predictors were included in each model unless a specific model selected them out during the training process. In total, [TO DO] kinds of models were used, with a pre-processing step (imputation, centering, scaling, and Box-Cox transformations) included in each iteration of model training under 10-fold cross-validation, repeated five times. Seeds were set in each instance to ensure reproducibility of the data, given the many assumptions and tuning parameters in our variety of models. All tuning parameters were selected using cross-validation in model training to find the optimal model that maximized AUC. Our models were:

- **Penalized logistic regression** (elastic net, binomial family, logit link), with objective function that includes loss and penalty given our high number of predictors, and tuning on α (mixing proportion) and λ (total penalization) for our final model
- **Generalized additive model (GAM)** (also binomial family, logit link), with potential inclusion of flexible nonlinearities for some predictors, performed in both `mgcv` for more flexibility and in `caret` for model comparison, and using general cross-validation to determine the smoothness of each operator $\hat{f}_j$ and to search over `GCV.Cp` given unknown scale parameter

- **Multivariate adaptive regression splines (MARS)** using `earth`, with tuning parameters (1) degree of features (number of possible hinge functions per parameter) and (2) number of terms (may not equal the number of predictors given the possibility of multiple hinge functions), including a stepwise model building procedure involving the addition of piecewise linear models / spline bases, followed by a pruning procedure
- **Linear discriminant analysis (LDA)**, which has no tuning parameters and assumes normally distributed features to classify by nearest centroid following data sphering and projection onto smaller subspaces, tends to work reasonably well with small n or well-separated classes, which doesn't appear true in our case and may make the model less robust
- **Random Forest**, [TO DO]
- **Boosting**, [TO DO]
- **Trees (CIT and CART)**, [TO DO]
- **Support Vector Machine (Linear and Radial Kernels)**, [TO DO – make sure to mention something about probabilities/ROC being poor metrics for SVM]
- **Neural Network**, [TO DO]

Optimal Tuning Parameters

Variable Importance

Resampling Results and Model Selection

Model Application to Test Data

Conclusion

TO DO

Still In Progress