P8106, Data Science II: Midterm Project

Zachary Katz (UNI: zak2132)

3/28/2022

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.

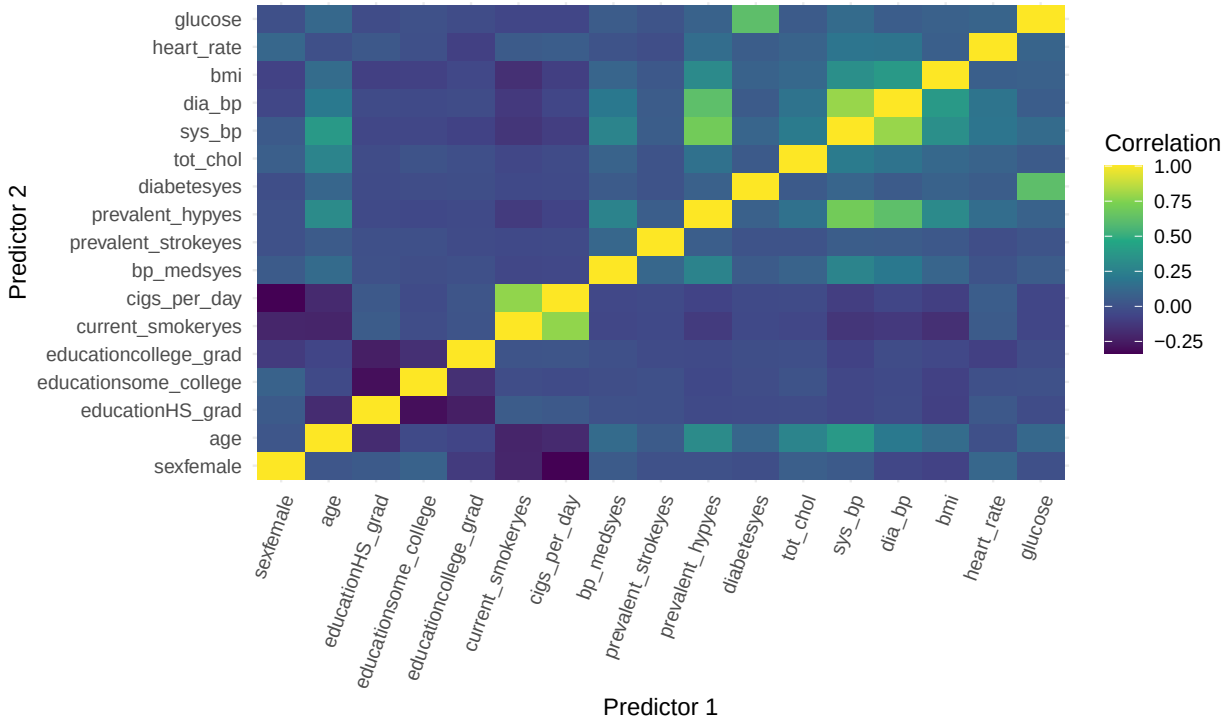
582 rows (13.7% of our observations) lacked at least one data point, with `glucose` accounting for the plurality of missing data (9.15% missing rate). We excluded observations with one or more missing data points from our study, but plan to include and impute missing data in a future iteration of our work. In total, this leaves 3,658 observations, of which 3,101 (84.8%) have CHD absent at 10 years, whereas 557 (15.2%) have CHD present – a notable class imbalance.

Exploratory Analysis & Visualization

Fig.1: Distributions of Predictors By Outcome Class

When we stratify the distributions of our continuous predictors by outcome status, we find the most substantial differences in median `age` and `sys_bp` 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 `prevalent_stroke`, `diabetes`, and `bp_meds`.

Fig.2: Heatmap of Correlations Between Predictors



We find 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.

Classification Modeling

Modeling Overview

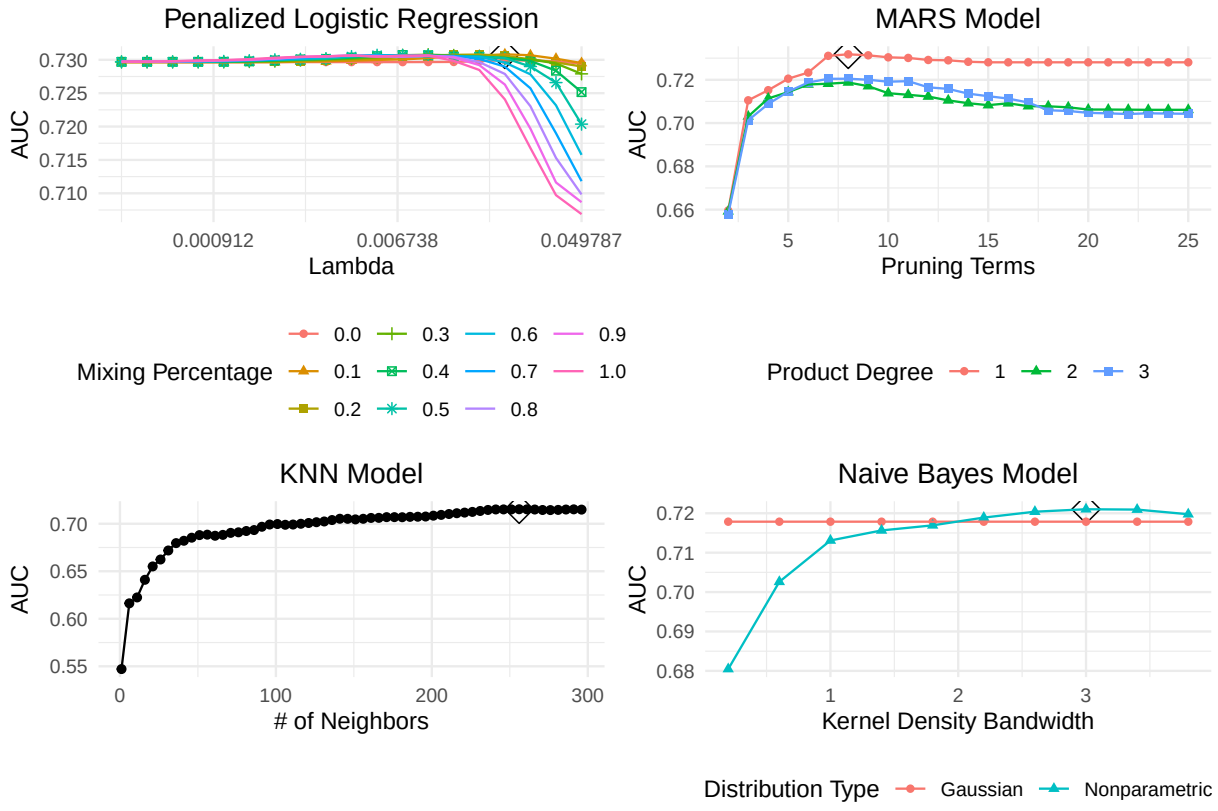
Initially, we attempted classification without rectifying class imbalance after splitting the data into 80% training set and 20% test set. Given lack of major collinearity and relatively low missingness, all available predictors were included in each model unless a specific model selected them out during the training process. In total, seven kinds of models were used, with a pre-processing step (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
- **K-Nearest Neighbors (KNN)** to predict class labels through majority vote among k neighbors, which is tuned using cross-validation, given known utility for class imbalance – but considered a "black box" model without clear relation between predictor and response
- **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
- **Quadratic discriminant analysis (QDA)**, like LDA, has no tuning parameters but permits flexible (quadratic) decision boundaries between classes, despite working better with well-separated response classes as well
- **Naive Bayes (NB)**, unlike LDA and QDA, assumes conditional independence of features in each class, which is generally more applicable for datasets with many qualitative and quantitative predictors; we include a Laplace correction given the possibility of test data points with feature values never before seen by the classifier, tuning our flexibility in the nonparametric case through class-validation on adjustments to kernel density bandwidths

Initial Modeling Results

Fig.3: Tuning Parameter Selection



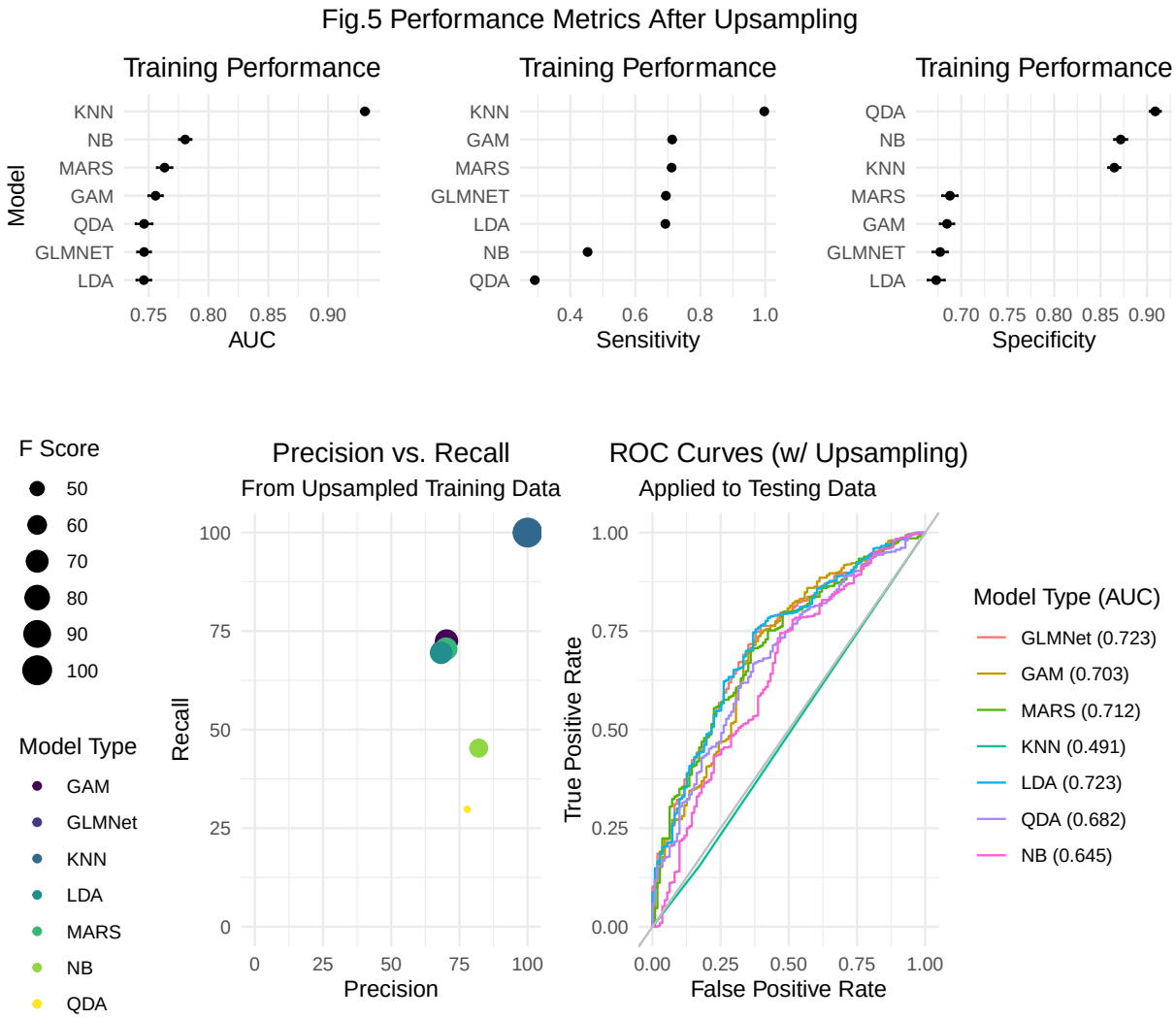
Our tuning parameters were selected by generalized cross-validation using only training data to maximize AUC. Our optimal elastic net model had $\alpha = 0.1$ (ratio between L1 and L2 penalty, which in our case is closer to ridge) and $\lambda = 0.0216$ (penalty strength parameter). For the MARS model, we select 1 product degree and 8 terms. In the KNN case, we choose to conduct majority voting with k = 256 neighbors, and for the NB nonparametric classifier with Laplace correction, we choose fL = 2, a higher kernel density bandwidth to promote flexibility. In addition, multiple models (e.g. elastic net and MARS) indicate that `age`, `sys_bp`, and `sex` are our three most important predictors for classifying response. Again, this is unsurprising given our exploratory data analysis and relatively better class separability for these features than for others. `cigs_per_day` also appears important to both models, whereas `glucose` seems more important to the MARS model than to the elastic net model.

Fig.4 Initial Performance Metrics

We find limited stratification by AUC, with the MARS model performing best (median AUC of 0.733), followed by elastic net (median AUC of 0.731) and GAM (median AUC of 0.727). The worst performer on this metric was QDA (median AUC of 0.699). QDA also has by far the highest sensitivity (17.8%) but lowest specificity (94.8%). Based on this training performance, and naive to any performance on testing data, we would choose the MARS model because it maximizes AUC and, second only to the QDA model (which has poorer AUC), also maximizes F-score and sensitivity (recall), which means that the test is better at

detecting the positives – quite important in the medical context. Normally, we would only apply our chosen model to the test data based on training performance, but out of interest, we checked the ROC curves for all of them. The MARS model has the second best AUC (0.722), while the QDA has the second-worst AUC. Generally, our ROCs look only "OK", and our 85% accuracy for most models is close to the No Information Rate, which isn't that insightful given the class imbalance.

Next Steps with Class Imbalance

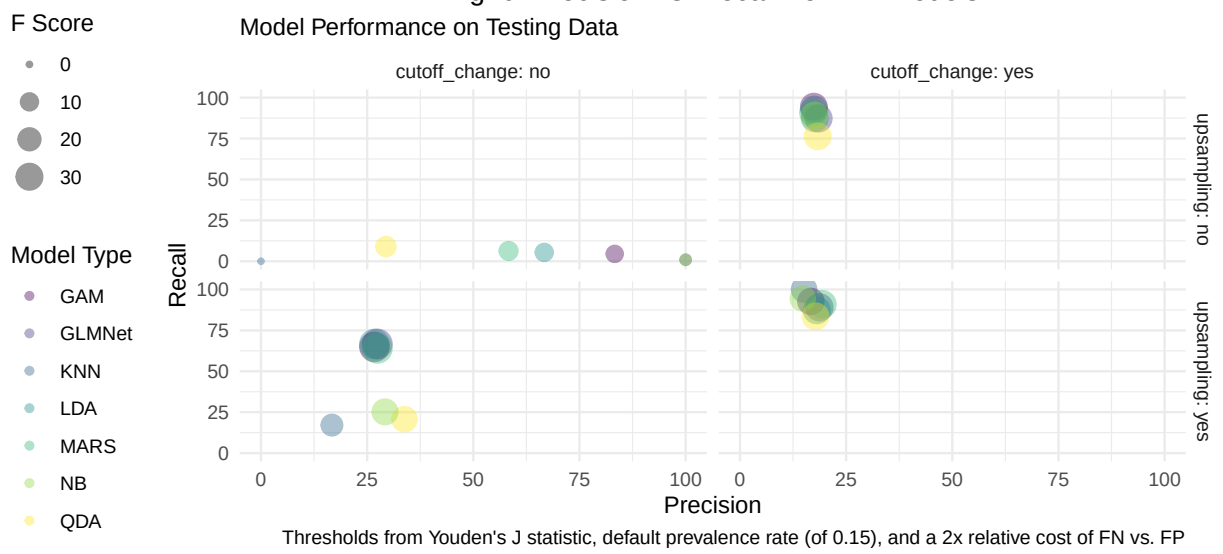


Given class imbalance, we first attempted to upsample from our `CHD_positive` training observations. Through upsampling, we improved our training model performance, with median KNN AUC reaching as high as 0.93 and an F-score of 100%. Purely based on these metrics, we may have chosen the KNN model as our final model, but from an application perspective, the lack of clear connection between predictors and outcome is problematic. In addition, once we looked at how our models performed on the test data with ROC curves, the KNN was indeed the worst — even worse than it was *without* upsampling, likely because of underfitting (as our k tuning parameter seems high compared to $\sqrt{n}$). Overall, while our model performance improved on the training data with upsampling, it did not improve on the test data.

We try one last approach, which is to vary the classification probability threshold, generally 0.5 by default. Because ROC curves are invariant to such transformation, we did this by reserving half of our test data for

an intermediate step, which we call "validation", before "true" testing. Once we've trained our models on the original 80% of observations, we then make predictions on our validation data set and use those predictions to optimize our new threshold before applying the new cutoff to our test prediction probabilities for both the upsampled training data and the original training data. We also conducted similar performance tests without changing the probability class cutoff to better ascertain impact of the methodology.

Fig. 6: Precision vs. Recall from All Models



Compared to baseline (no upsampling or cutoff change), upsampling led to greater precision improvement, cutoff changing led to greater recall, and F-scores were maximized with upsampling but no cutoff change - especially for elastic net (F-score of 38.8%) and MARS model (F-score of 38.4%) — when applied to the test data. However, it's poor practice to select a model based on test performance. Based purely on training data performance, we would have gone with our original MARS model over the upsampled KNN model because of the "black box" nature of KNN, and because its anomalous training AUC would have cautioned us of lack of fit.

Limitations

First, we do not include any method of imputation. In addition, because our data has imbalanced classes, we upsampled without having learned the ideal way to do so. We exclude interaction terms in this analysis, although there may very well be some effect measure modification. Finally, we look forward to building on this work through alternative machine learning models, including random forest, SVM, and boosting.

Conclusion

Overall, our modeling tells us that the most important determinants of heart disease tend to be age, blood pressure, glucose, sex, and the number of cigarettes smoked per day, which jibes with our exploratory analysis. That said, there may be other covariates, such as race and income, that may have stronger explanatory power of our `ten_year_chd` outcome. Given the predictors at our disposal, our attempts to find a model with an AUC over 0.80 did not prevail without upsampling, barring a likely-underfit KNN model. However, even without resampling or changing the probability cutoff for classification as a `CHD_present`, we had a MARS model with a reasonably high AUC and F-score. In particular, this is a critical measure in our case because the penalty for a false negative should be higher than for a false positive. However, a key limitation of the MARS method is the fact that it operates in a stepwise greedy manner, and so a future solution may involve considering all possible hinge functions simultaneously before running penalized regression, with the tradeoff of high computational expensive.