P8106 Midterm: Framingham Heart Study

Tucker Morgan - tlm2152

3/22/2022

Contents

Introduction

In this report, I will examine a data set from the Framingham Heart Study (linked). I will use various machine learning techniques in an attempt to predict the ten-year risk of coronary heart disease (CHD), `ten_year_chd`, a binary outcome, based on a set of 15 predictors. After removing NA values, the full data set has 3658 observations. Luckily, the proportion of observations with NA values and `ten_year_chd` = 1 (positive) is similar to the proportion of observations with NA values and `ten_year_chd` = 0 (negative), 0.1537 for positive versus 0.1518 for negative. Several variables require re-coding and factorizing before work begins, including `education`, `sex`, `current_smoker`, `bp_meds`, `prevalent_stroke`, `prevalent_hyp`, `diabetes`, and `ten_year_chd`.

The data is split into 80% training data with 2927 observations and 20% testing data with 731 observations. The training data will be used in exploratory analysis and for model training with the testing data being used after model selection to evaluate test performance.

Exploratory Analysis

First, looking at Figure 1 in the Appendix, we see the collinearities between the continuous predictors in the data set, the highest being just 0.78 between systolic (`sys_bp`) and diastolic (`dia_bp`) blood pressure. While this is notable, it is not so high as to be concerning or to warrant removal of predictors. We also see in Figure 2 there are not major differences between CHD-present (`ten_year_chd` = 1) and CHD-absent individuals for many of the continuous predictors. The predictors `dia_bp` and `sys_bp` tend to be slightly higher for CHD present individuals, but by a very small margin. The only continuous predictor with clear differences is `age` - CHD-present individuals tend to be older than CHD-absent individuals.

In Table 1 of the Appendix, we see the percentage of CHD-present individuals is higher for males, individuals taking blood pressure medication (`bp_meds`), individuals who have experienced stroke (`stroke` or `prevalent_stroke`), individuals who have hypertension (`hyp` or `prevalent_hyp`), and individuals with diabetes (`diabetes`). There might be a relationship between `education` and `ten_year_chd` - we see a higher percentage of CHD positive individuals with just some high school education (Table 2). However, this is not as clear.

Models

I am investigating four models as candidates for predicting `ten_year_chd`. Each model is trained on all predictors in the training data and assessed using cross-validation area-under-curve (AUC). Each of the models has different strengths, weaknesses, tuning parameters, and assumptions. Because we did not see concerning colinearities or relationships in exploratory analysis, I am not removing any predictors prior to fitting. However, I am scaling and centering the predictors in each model.

The first model is a **generalized linear model (GLM)** logistic regression with no tuning parameters or variable selection. This will serve as an interesting baseline to compare against the more complicated models to follow. As the name implies, logistic regression is a generalization of linear regression. In linear regression, we assume a linear relationship between some outcome, Y, and predictors $X_1, ..., X_n$. The same is true for logistic regression, except the outcome is $\log \frac{\pi}{1-\pi}$ where π is a probability of belonging to a class - here `ten_year_chd` = 1. An advantage of GLM is relatively easy interpretability, however it has limited flexibility and may not accurately reflect the truth if non-linear relationships are in play. The fitted coefficients for the GLM can be seen in Table 3. We see that `sexfemale`, `age`, `sys_bp`, `cigs_per_day`, and `glucose` have the strongest relationships with our outcome. It will be interesting to see how this compares to other models.

The second model is an **elastic net model** with a mixing parameter for least absolute shrinkage and selection operator (LASSO) and ridge regression penalties. This is another logistic model with the same assumptions

as in GLM. However, there are two tuning parameters, α and λ. Parameter λ indicates a penalty term that limits the number and/or magnitude of predictor coefficients in a model. There are two types of penalties - the ℓ_1-norm associated with LASSO and the ℓ_2-norm associated with ridge. Parameter α determines the mixture of these two penalties with $\alpha = 0$ being ridge regression and $\alpha = 1$ being LASSO. The optimal values for these tuning parameters are chosen via maximizing cross-validation AUC. The tuning parameter results can be seen in Figure 3, with an optimal mix of parameters is $\alpha = 1$ and $\lambda = 0.0093$ resulting in the largest AUC. The selected variables can be seen in Table 4, and these align somewhat with the strongest relationships observed with GLM.

The third model is a **multivariate adaptive regression splines (MARS) model**, which can incorporate non-linear features through products of hinge functions. It is typically well-suited for high-dimensional problems. There are again two tuning parameters here, the degree of interaction and the number of retained terms. A higher degree allows for higher order interactions of hinge functions. The tuning parameter results can be seen in Figure 4 with a degree of 1 and number of terms equal to 7 being optimal. The selected variables and associated hinge functions can be seen in Table 5. Again, we see some of the same predictors featured in MARS as in the previous models, namely `age`, `sys_bp`, `sexfemale`, `glucose`, and `cigs_per_day`. In the variable importance plot (VIP) in Figure 5, we see which variables were found to be most influential in the model building process.

The fourth model is a **naive Bayes (NB) model**, which assumes all predictors are independent within each class and can use Gaussian or non-parametric distributions of predictors. The tuning parameters in the NB model are distribution type and bandwidth adjustment for the kernel density, which adds a small number to the counts for each feature to prevent non-zero probabilities in non-parametric estimations. The main weakness of NB is that the assumption of independent features is often incorrect. For example, in the Framingham Heart Study data set, `sys_bp` and `dia_bp` are very likely not independent measures since both measure blood pressure. The tuning parameter results for NB can be seen in Figure 6, where we see a non-parametric distribution and a bandwidth adjustment of 2.6 maximize the AUC (labeled "ROC").

Table 6 shows the results from cross-validation assessment of the four models. Based on these results, we select the elastic net model to be used for predictions with the testing data because it has the highest mean AUC. The final ROC curve and AUC for the elastic net against the test data can be seen in Figure 7, and a confusion matrix of predictions can be seen in Table 7.

Conclusions

The final model has an AUC of 0.7365, which is somewhat acceptable, but not as close to 1 (perfect prediction) as we would like. Looking at Table 7, we see the model greatly underestimates the number of positive `ten_year_chd` instances. The kappa value, which we would like to be positive and close to 1, is only 0.0421, and the sensitivity is only 0.027 meaning the model does not perform well in correctly identifying positive cases. The results in Table 7 are obtained using a decision boundary of $p = 0.5$, and adjusting this threshold can correctly classify a greater number of positive individuals as seen in Table 8. However, this also results in more false positives, a potentially distressing diagnosis with CHD, and this increases the overall misclassification error of the model from 0.1491 to 0.2367. This shows that changing the decision boundary will not result in the model being perfect.

At the $p = 0.5$ decision boundary, the model is predicting almost every observation to be negative, and this is correct most of the time due to the class imbalance between `CHD_present` and `CHD_absent`. Overall, I was hoping for better model performance, however I am interested to see if improvements could be made in future work by adjusting for the class imbalance and imputing missing data to obtain more information from the data set.

Appendix

Figures

Figure 1: Continuous Predictor Correlations

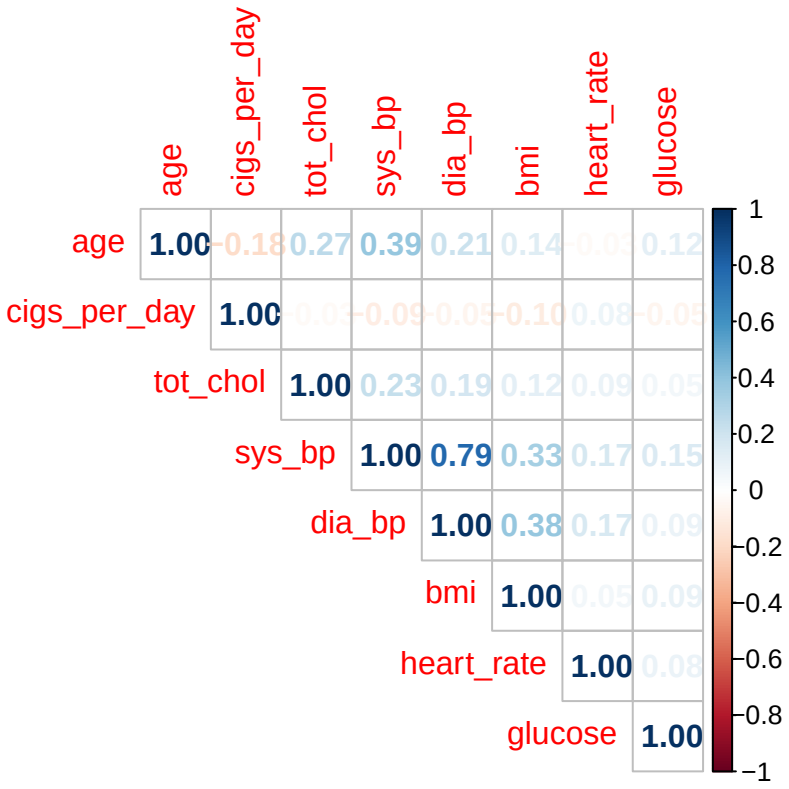


Figure 2: Continuous Predictor Feature Plot

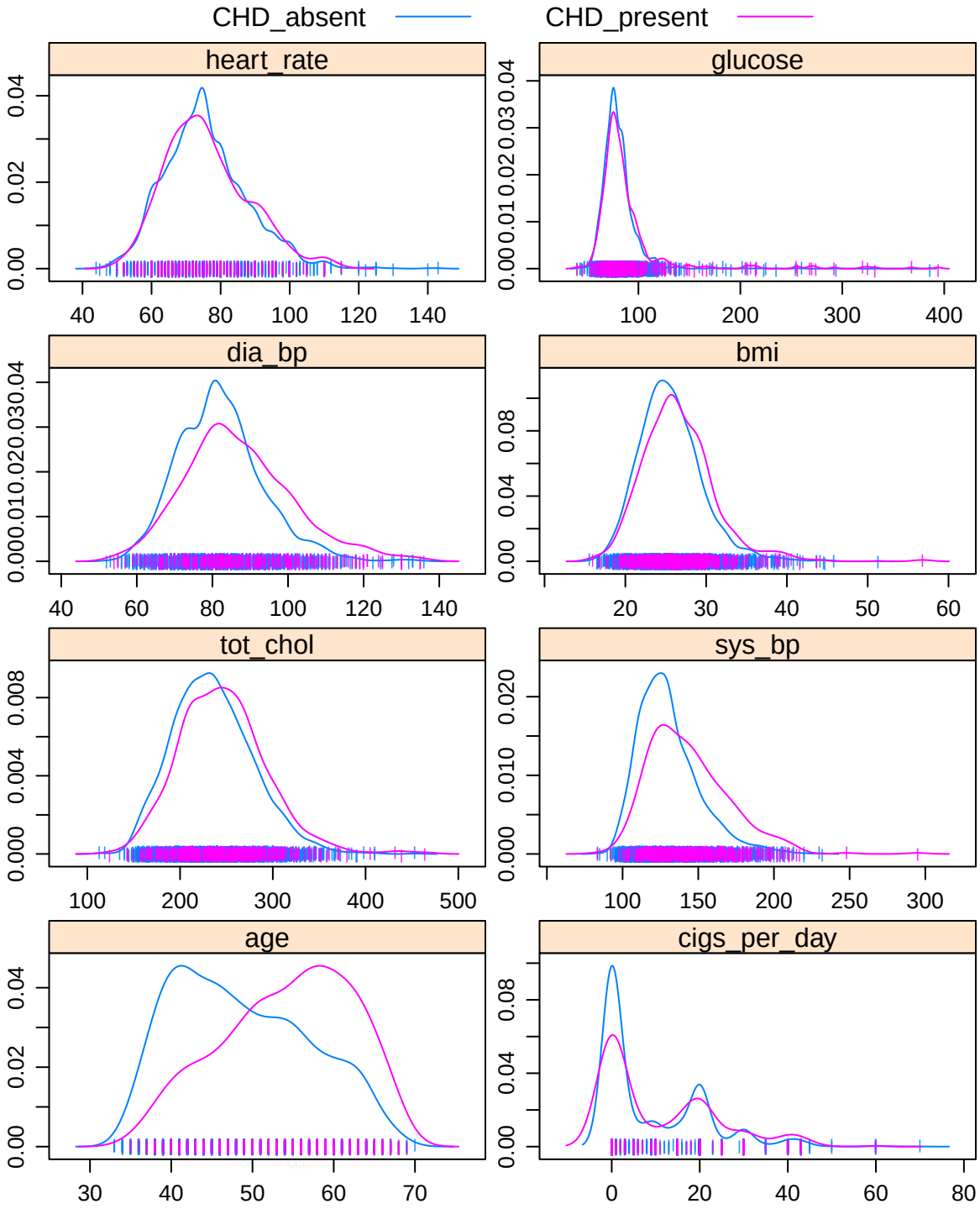


Figure 3: Elastic Net Tuning Parameters

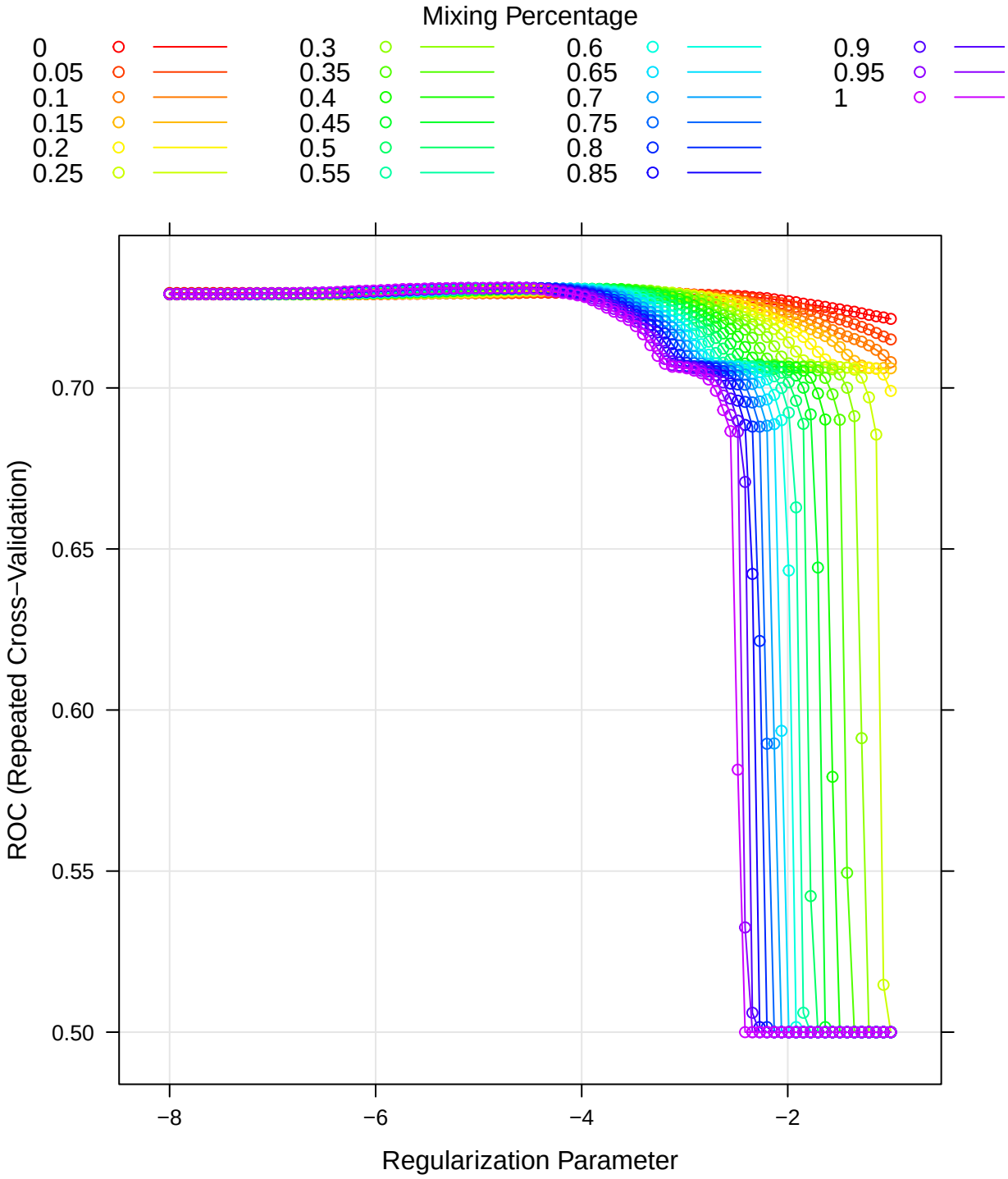


Figure 4: MARS Tuning Parameters

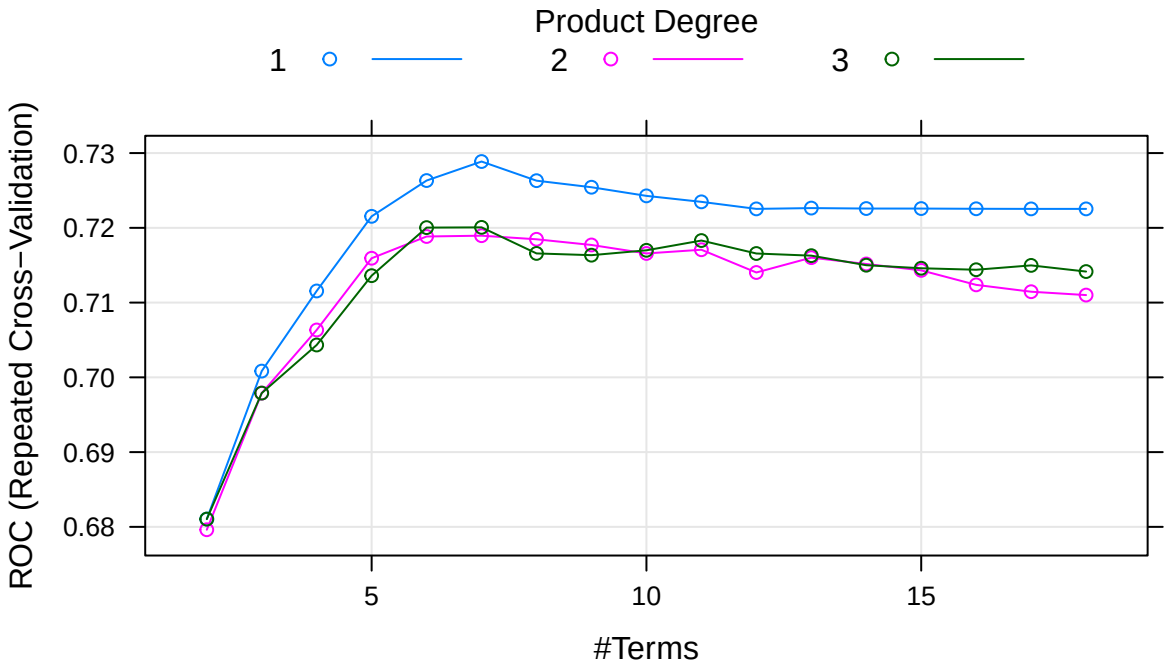


Figure 5: MARS Variable Importance Plot

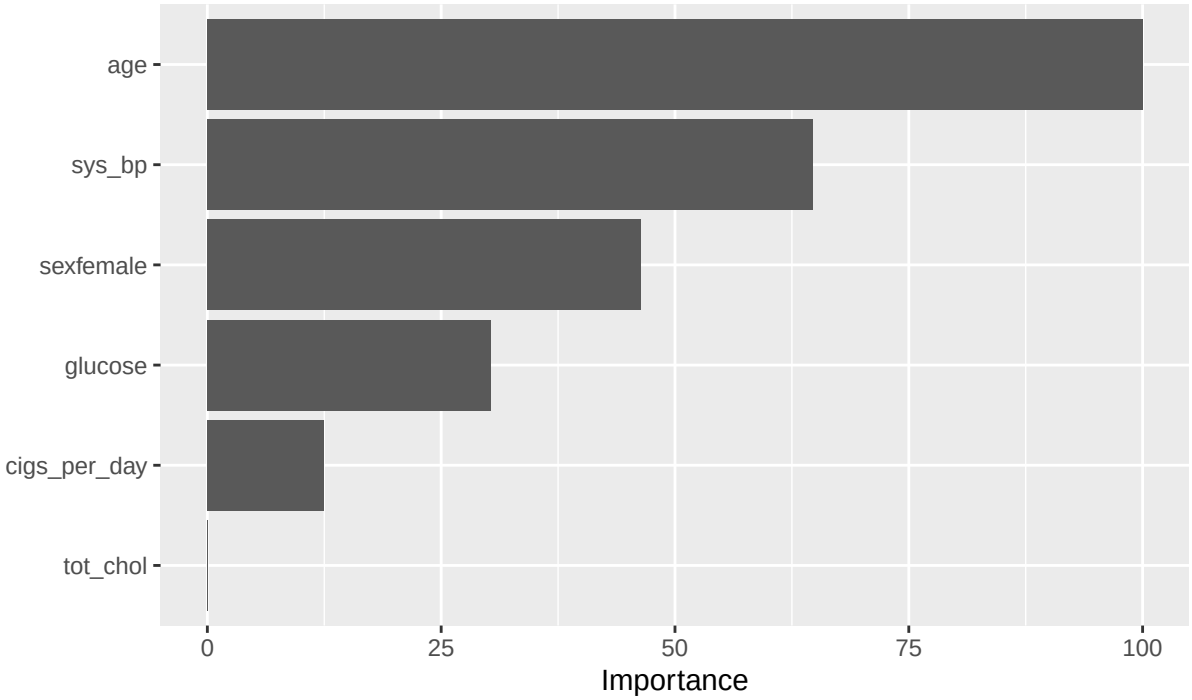


Figure 6: Naive Bayes Tuning Parameters

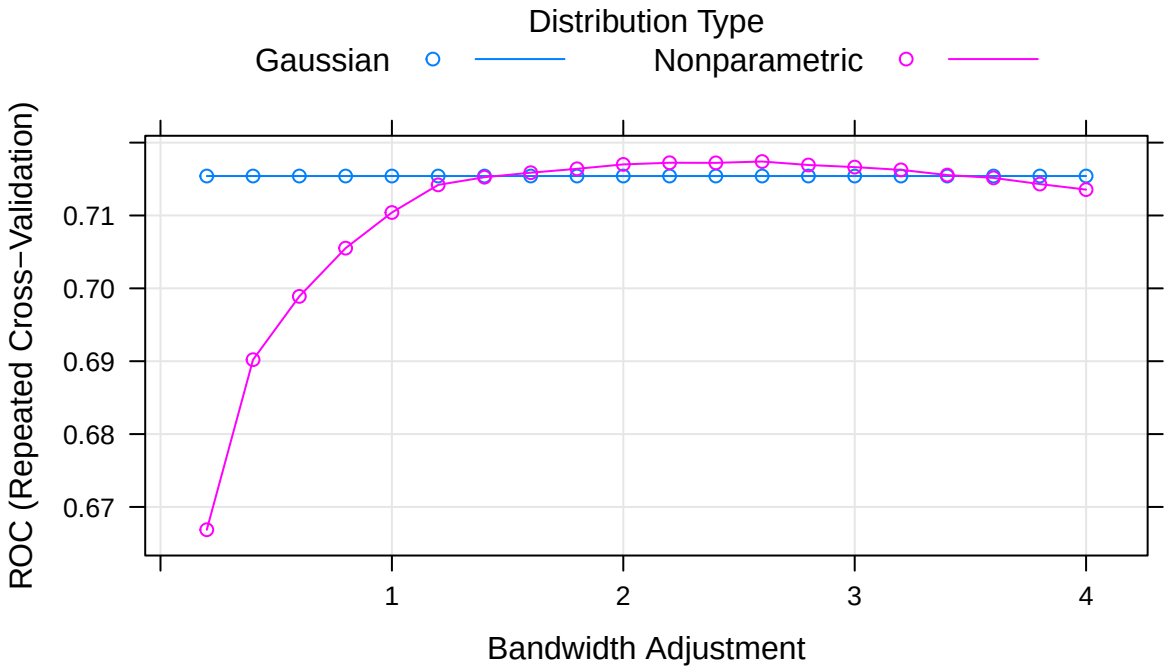
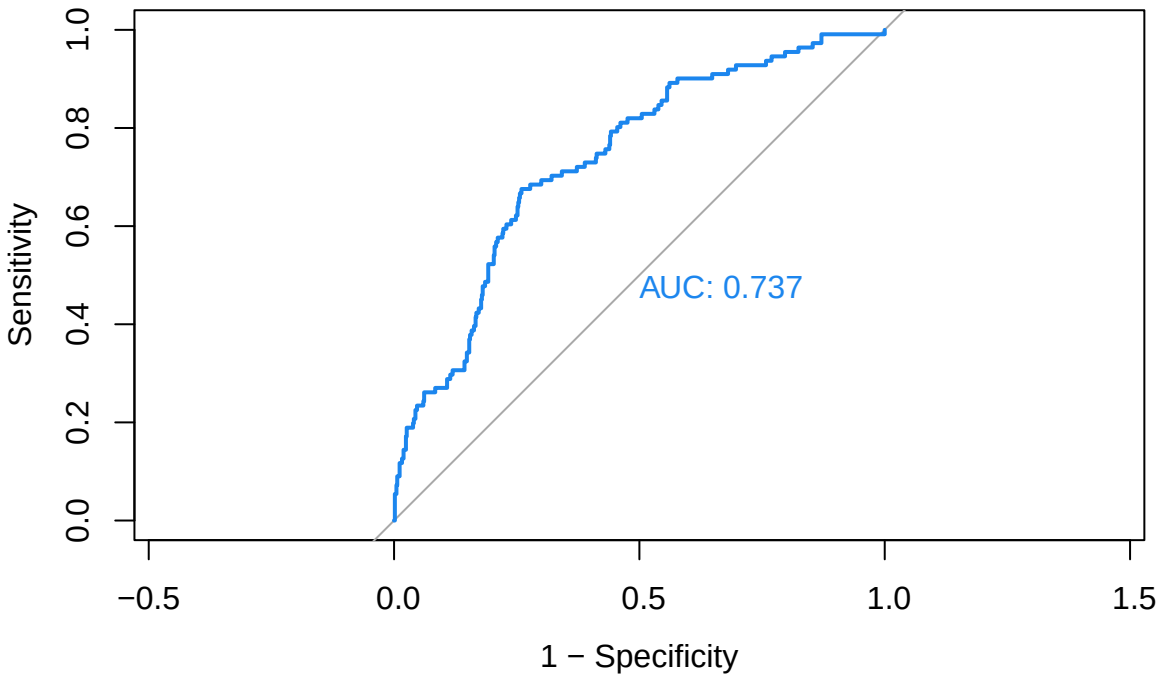


Figure 7: Elastic Net ROC Curve



Tables

Table 1: Ten-Year CHD % Positive for Binary Variables

sex	chd_1	cur_smoke	chd_2	bp_meds	chd_3	stroke	chd_4	hyp	chd_5	diabetes	chd_6
male	0.20	no	0.15	no	0.15	no	0.15	no	0.11	no	0.15
female	0.12	yes	0.16	yes	0.31	yes	0.44	yes	0.25	yes	0.37

Table 2: Ten-Year CHD % Positive for Education Levels

education	chd_perc
some_HS	0.19
HS_grad	0.12
some_college	0.12
college_grad	0.14

Table 3: GLM Coefficients

	x
(Intercept)	-1.9988268
sexfemale	-0.2994759
age	0.5329661
educationHS_grad	-0.0742298
educationsome_college	-0.0641615
educationcollege_grad	-0.0160658
current_smokeryes	0.0437312
cigs_per_day	0.1998258
bp_medsyes	0.0227162
prevalent_strokeyes	0.0780255
prevalent_hypyes	0.1222129
diabetesyes	-0.0035051
tot_chol	0.1057025
sys_bp	0.3137888
dia_bp	-0.0616929
bmi	0.0407147
heart_rate	-0.0321289
glucose	0.1965919

Table 4: Elastic Net Coefficients

	s1
(Intercept)	-1.9057558
sexfemale	-0.2277163
age	0.4774582
cigs_per_day	0.1434178
prevalent_strokeyes	0.0391234
prevalent_hypyes	0.0817260
tot_chol	0.0286909
sys_bp	0.2603499
glucose	0.1491647

Table 5: MARS Coefficients

	x
(Intercept)	-2.0052943
h(age- -0.538629)	0.6906015
h(sys_bp-0.076422)	0.4821964
sexfemale	-0.2946299
h(glucose-0.309344)	0.2592104
h(2.8538-cigs_per_day)	-0.2332780
h(tot_chol-3.07093)	1.5425476

Table 6: Cross-Validation AUC Values

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
GLM	0.6643463	0.7076649	0.7274194	0.7288203	0.7483672	0.8333333	0
NET	0.6617581	0.7062307	0.7350939	0.7311366	0.7496237	0.8215054	0
MARS	0.6520303	0.7060820	0.7304884	0.7288690	0.7489775	0.8400090	0
NB	0.6286657	0.6955665	0.7153592	0.7174034	0.7429206	0.7923835	0

Table 7: Elastic Net Confusion Matrix

	CHD_absent	CHD_present
CHD_absent	619	108
CHD_present	1	3

Table 8: Adjusted Elastic Net Confusion Matrix - p = 0.2

	CHD_absent	CHD_present
CHD_absent	501	54
CHD_present	119	57