Parametric vs. Non-parametric Methods in Machine Learning

Hun Lee (sl4836), 3/27/2022

Introduction

The goal of the project is to examine whether non-parametric prediction methods in machine learning can perform better than parametric methods when the data set does not follow the assumptions of parametric prediction models.

Data description and dictionary (binary variable - 1: yes, 0: No)

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects.

sex : The gender of the observations (male & female).
age : Age at the time of medical examination in years.
education : Some high school (1), high school/GED (2), some college/vocational school (3), college (4)
currentSmoker: Current cigarette smoking at the time of examinations
cigsPerDay: Number of cigarettes smoked each day
BPmeds: Use of Anti-hypertensive medication at exam
prevalentStroke: Prevalent Stroke (0 = free of disease)
prevalentHyp: Prevalent Hypertensive. Subject was defined as hypertensive if treated
diabetes: Diabetic according to criteria of first exam treated
totChol: Total cholesterol (mg/dL)
sysBP: Systolic Blood Pressure (mmHg)
diaBP: Diastolic blood pressure (mmHg)
BMI: Body Mass Index, weight (kg)/height (m)^2
heartRate: Heart rate (beats/minute)
glucose: Blood glucose level (mg/dL)
TenYearCHD (outcome variable): The 10 year risk of coronary heart disease(CHD).

Before diving into exploratory analysis and model training, missing data is omitted in this project in order to compare the model performance in two different scenarios, omitting missing data vs. imputing missing data, which will be done and reported in the upcoming final project. Because the original data set was already quite clean, not much was needed to be done for data cleaning other than omitting missing data, cleaning variable names, turning categorical variables into factors, and re-leveling the outcome variable to predict whether one would have chronic heart disease. Lastly, to measure the performance of differnet models with testing data set, the data set is divided into training and testing data set with the ratio of 8 to 2.

Exploratory analysis/Parametric Model Assumption Check

From Fig.1, we can check whether each class has equal variance/covariance in the predictors among two CHD classes from covariance ellipses and box plots. It is to be observed the variance is unequal among two CHD groups for *age*, *sys_bp* and *cigs_per_day* variables. It is also to be observed that the covariance ellipse for CHD group is generally much wider than NoCHD group. The density plots in Fig.1 also shows that the continuous predictors given the outcome variable, $f(X|Y = y)$, are fairly normally distributed for *heart_rate* variable, but not to be normally distributed for *age*, *cigs_per_day*, *glucose*, and *sys_bp* variables. Another thing to note is that two CHD groups are not well dispersed by continuous predictors except *sys_bp* and *age* variables and hence they are expected to play more important roles in predicting the status of chronic heart disease than the other continuous variables. In light of the above exploratory analysis, the assumptions of normality and equality variance-covariance matrices are not met with our data.

Additionally, checking of Fig.2 for the assumptions of a logistic regression lets us see that the data does not satisfy the assumptions of homogeneity of variance and the normality of residuals. An important assumption of logistic regression is that the errors (residuals) of the model are approximately normally distributed because the model has a nonlinear transformation of the predicted values, so the degree to which observed values deviate from the predicted values is expected to vary across a range of values, with most residuals being near 0 and fewer residuals deviating far from the predicted line. Based on the these analyses, we expect parametric methods, such as Logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), or parametric (Gaussian) Naive Bayes not to be the best methods for predicting whether one would have chronic heart disease (CHD) with our data set.

Regression methods that uses regularization, such as least absolute shrinkage and selection operator (Lasso), are expected to be better methods considering it does not require assumptions like normality or equal variance-covariance matrices and our data set does not have influential outliers and there are 15 predictors (high dimensionality) among which are continuous and not highly correlated. However, Lasso is also a parametric method and hence we are not confident in that it is going to perform better than non-parametric methods.

Among categorical variables, *sex* and *education*, and *prevalent_hyp*, have a fair amount of difference of proportions among two CHD classes; hence, they may play more important roles in predicting whether one would have chronic heart disease than the rest of the categorical variables.

Models & Results

To fit and train models for machine learning classifiers, all 15 predictors are used as predictors in all parametric and non-parametric models and they are all centered and scaled through data pre-processing. The parametric models for machine learning classifiers with the training set are Logistic regression, Lasso regression, parametric Naive Bayes, and LDA method and these models require distributional and parametric assumptions. For logistic regression, the assumptions of homogeneity of variance, the normality of residuals, no influential outliers, and no multicollinearity are required. For LDA method, the assumptions of equal variance-covariance matrices and conditional Normality $f(X|Y = y) \sim N$ are required. For Gaussian (parametric) Naive Bayes, the predictors associated with each response class are expected to be normally distributed with the independence assumptions between predictors. Lastly, for Lasso method, there should be no influential observations.

Unfortunately, as aforementioned, most of the assumptions of the parametric models are not met with our data set; hence, this project includes non-parametric models in order to check if indeed non-parametric models can be better methods than parametric models when the assumptions of the parametric models are violated. The non-parametric models for machine learning classifiers with the training set are multivariate adaptive regression spline (MARS), supervised principal component analysis (PCA), and non-parametric Naive Bayes method. MARS can be used to create a set of hinge functions that result in discriminant functions that are nonlinear combinations of the original predictors for classification. For supervised PCA, the eigenvectors of covariance matrix are used to find principal components and use them for dimensionality reduction. Among 15 components, 12 components cover about 95% of the variance, the first component explains 21.6% and the second 12.6% (Fig.4). Since 12 components explain the most of the variance, the supervised PCA model is fitted with these 12 components and turns out 12 components are the best number of variables with the highest estimated mean of AUC.

For the models that use tuning parameters, such as Lasso, MARS, and Naive Bayes, the best tuning parameter is selected such that the highest area under the ROC curve (AUC), known as the measure of the ability of a classifier to distinguish between classes (separabiltiy), is obtained from 10-fold cross validation (repeated cross-validation did not give better result and was more time-consuming and hence 10-fold cross validation method was chosen). Lasso model chooses lambda value 0.003506 as the best tuning parameter and Naive Bayes with non-parametric model with 2 flow (Laplace Smoother) as the best tuning parameter (Fig.4). In fact, all of 0, 1, and 2 flow are the best tuning parameter because they give the same result. For MARS model, 6 prunes (nprune) are selected as the best tuning parameter and five predictors are selected. Among the five predictors, sys_bp takes the highest importance role, followed by age, glucose, cigs_per day, and sex (female). In fact, sys_bp and age were expected to play more significant roles in predicting chronic heart disease than the other variables in the previous part. Lastly, the product degree of the MARS model is 1, implying that there is no interaction in hinge functions.

Note that Area under the ROC curve (AUC) is the metric used to compare the performance of the models and Fig.5 shows the cross-validated result of all the models with the training data set. Two non-parametric models, MARS and supervised PCA (in order), perform better than all the parametric models based on the result of the estimated mean of AUC from 10-fold cross validation, followed by Lasso, Logistic regression, LDA, and Naive Bayes. For the model performance with the testing data set, supervised PCA has the highest AUC, followed by Lasso, Logistic regression, MARS, LDA, and Naive Bayes (Fig.6). Considering that many assumptions of parametric models are violated with the data set, these results make sense. It is also not surprising that the cross-validated Naive Bayes method chooses the best tuning parameter from a nonparametric method with higher estimated mean value of AUC than Gaussian method (Fig.4).

Besides the violation of the assumptions of the parametric models, one of the limitations in this project is the type and the number of predictors to predict whether one would have chronic heart disease. I believe the prediction result would have been better with more number of predictors associated with chronic heart disease. Another limitation is the imbalanced number of the response variable class. It is recommended to have at least one thousand observation in each class, but we only have 644 observation in CHD group and 454 observations in training CHD group. If we were to have more observations in CHD group, the prediction accuracy would be expected to be higher. Though the predictors in the data set are not the best variables, MARS, PCA, and Lasso regression are flexible enough to capture the underlying truth with appropriate tuning parameters and data pre-processing.

Conclusion

For the final model selection, MARS model will be chosen for the prediction of chronic heart disease because it has the highest estimated mean of AUC from 10-fold cross validation using the training set. On the one hand, there is little discernible difference in the estimated mean of AUC between MARS (0.7350) and supervised PCA(0.7330); hence, supervised PCA model can be also considered as the final model for the prediction of chronic heart disease.

Given that non-parametric methods in machine learning show the better performance over parametric methods, it is to be concluded that for non-parametric models have more flexibility by taking a large number of various functional forms and hence may have more prediction power by not being regulated by the parametric (distributional) assumptions unlike parametric methods, especially under the circumstance where those assumptions are violated. This conclusion underpinned from the Fig.8 that compares two model (LDA vs PCA) plots. The plot fitted by the supervised PCA model shows that principal component variables have two different classes of response variable observations more well dispersed and separated with more flexibility in the model compared to the plot fitted by the LDA model which shows that linear discriminant variable does not successfully have two different classes of response variable observations well dispersed (two classes are mostly overlapped). Considering that key assumptions of LDA method are violated, this result is not surprising. Thus, it is also to be concluded that parametric methods in machine learning are constrained to the specified functional forms and parametric assumptions and hence may be more prone to different features and types of data set than non-parametric methods. Among parametric models, Lasso regression performs the best. This outcome is reasonable considering Lasso regression is not constrained to much parametric assumptions, such as normality, equal covariance matrices, or independence between predictors. Thus, it is recommended to use non-parametric methods or Lasso regression method in machine learning prediction when data does not meet those three assumptions.

Last but not least, the work of this project finds systolic blood pressure, age, and glucose to be important factors in predicting whether one would have chronic heart disease or not. However, it should be noted that it is likely that there are other important variables for chronic heart disease prediction which this data set does not have; hence more prediction analyses need be to done with more number of potential variables, larger number of data size, and possibly data set with more balanced classes of the outcome variable in order to make a stronger and more confident conclusion.

Reference

Datasciencediving. (2017, November 3). Principal component analysis in R. Data Science Diving. https://datasciencediving.wordpress.com/2017/10/05/principal-component-analysis-in-r/

Diagnostics for logistic regression - web.pdx.edu. (n.d.). https://web.pdx.edu/~newsomj/cdaclass/ho_diagnostics.pdf

Divyariyer. (2021, January 19). Heart disease prediction - framingham casestudy. Kaggle.Divyariyer. (2021, January 19). Heart disease prediction - framingham casestudy. Kaggle. https://www.kaggle.com/code/divyariyer/heart-disease-prediction-framingham-casestudy/data

Framingham Heart Study - biolincc.nhlbi.nih.gov. (n.d.). https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation_2021a.pdf?link_time=2022-03-25_09:22:37.141675

Appendix

Checking the Assumptions of Parametric Models & Exploratory Analysis

Fig.1 Checking the Assumption of Conditional Normality $f(X|Y = y) \sim N$ and Equal Variance-Covariance

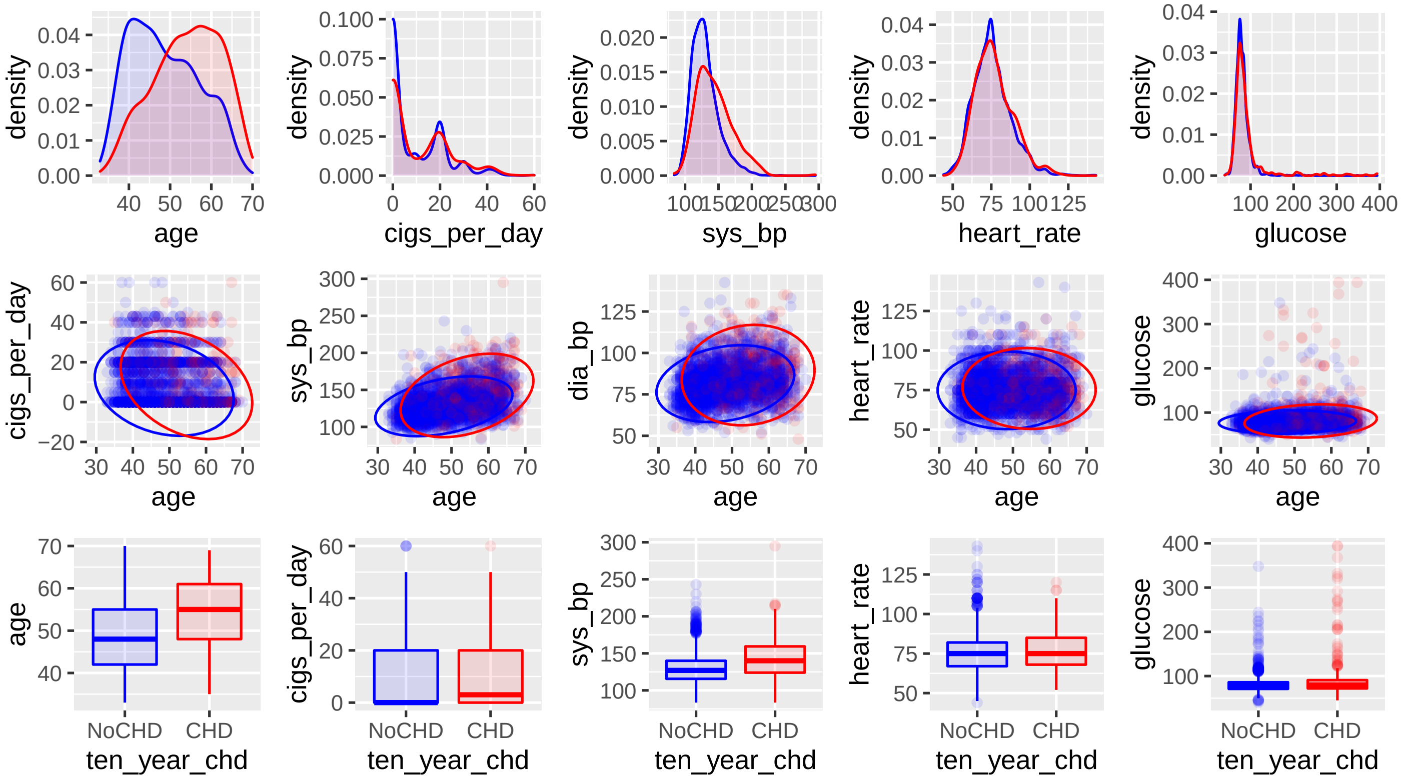


Fig.2 Checking the Assumptions of Homogeneity of Variance and Normality of Residuals

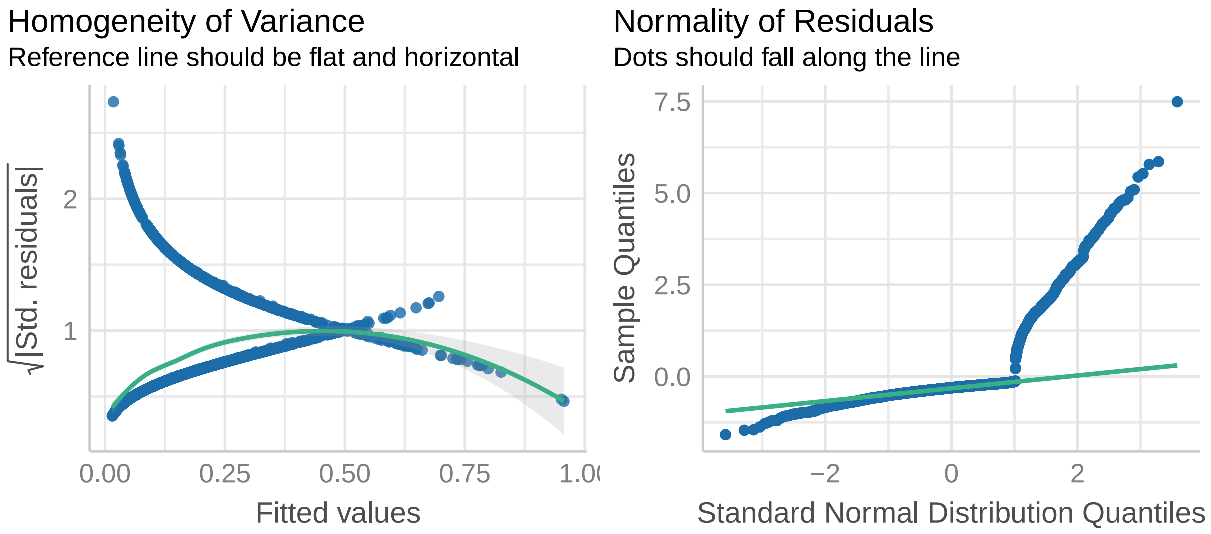


Fig.3 Model Tuning Parameter

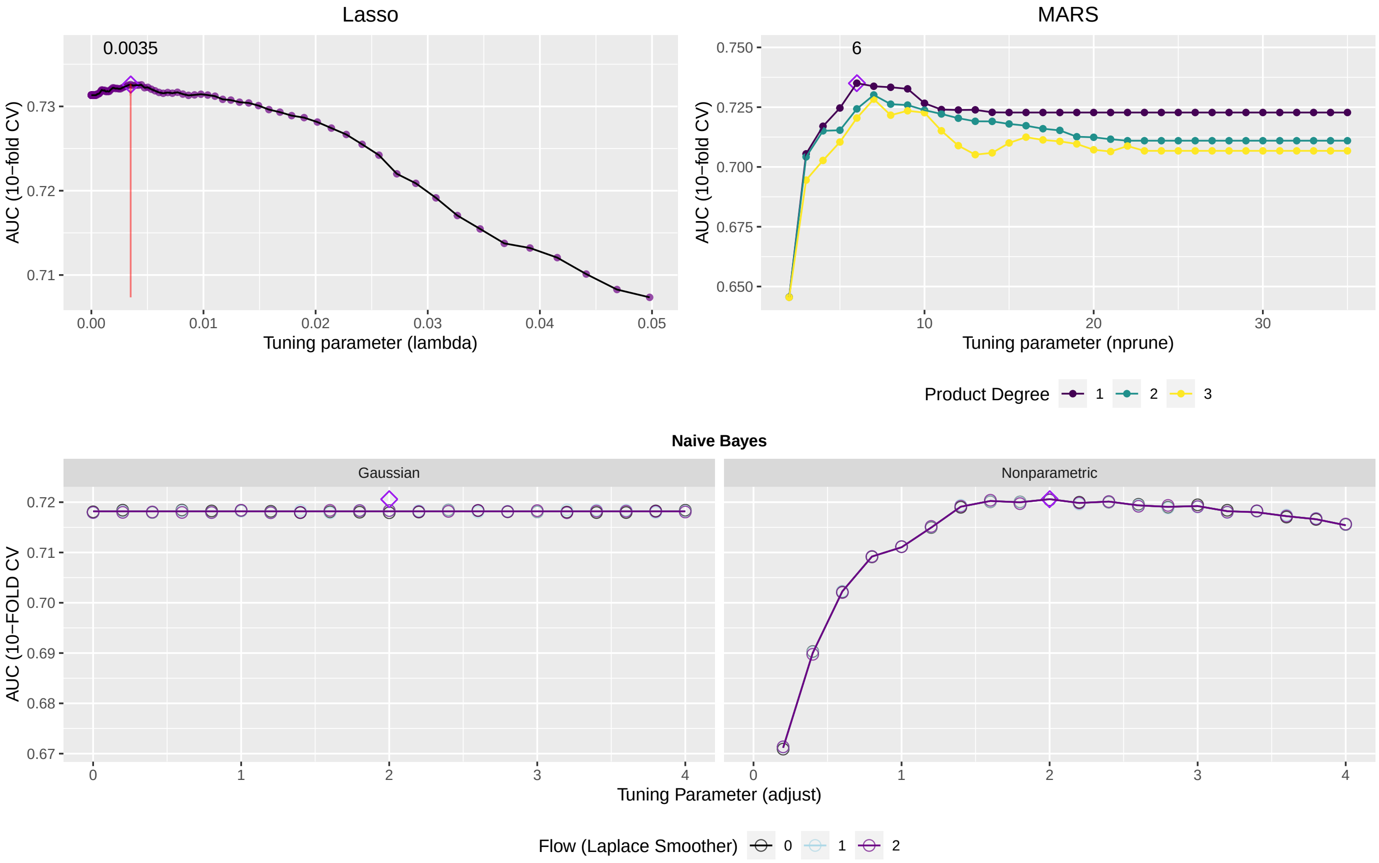


Fig.4 LDA Plot vs. PCA Plot

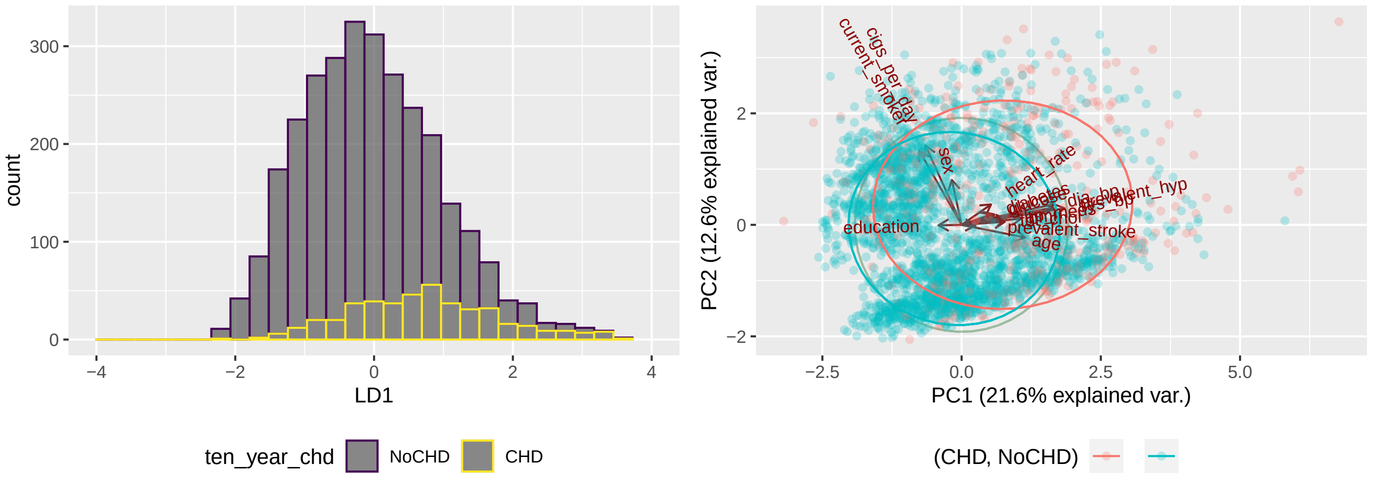
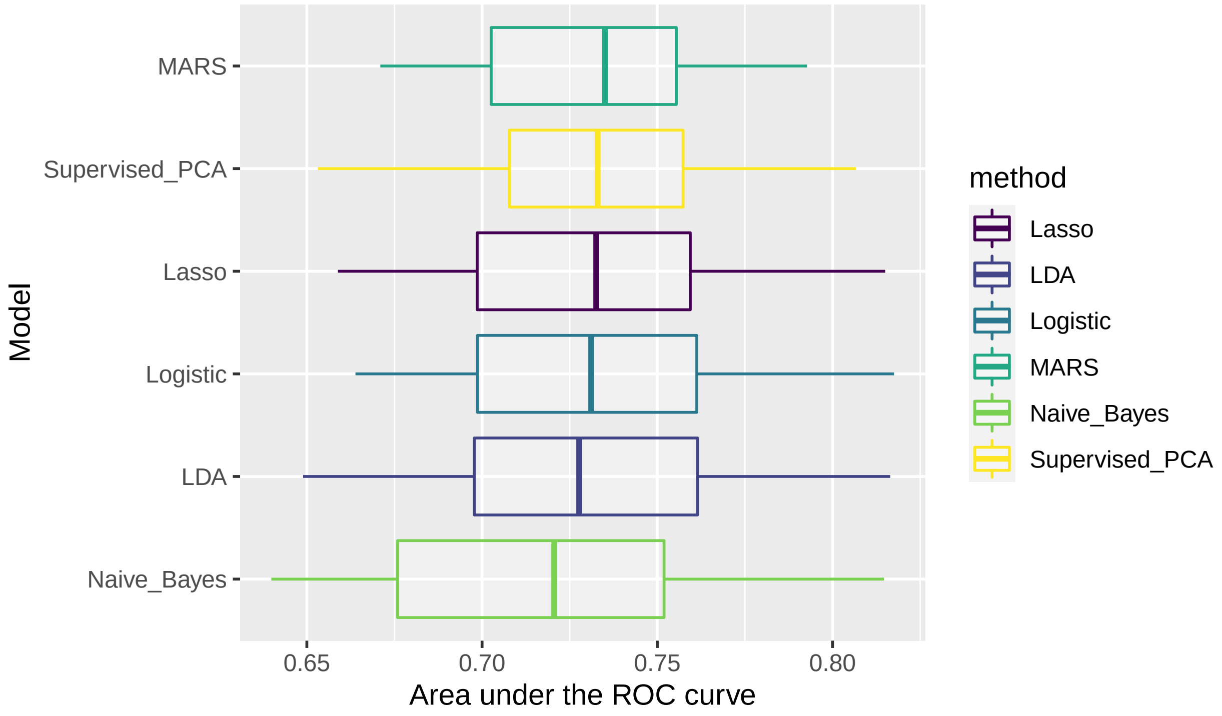


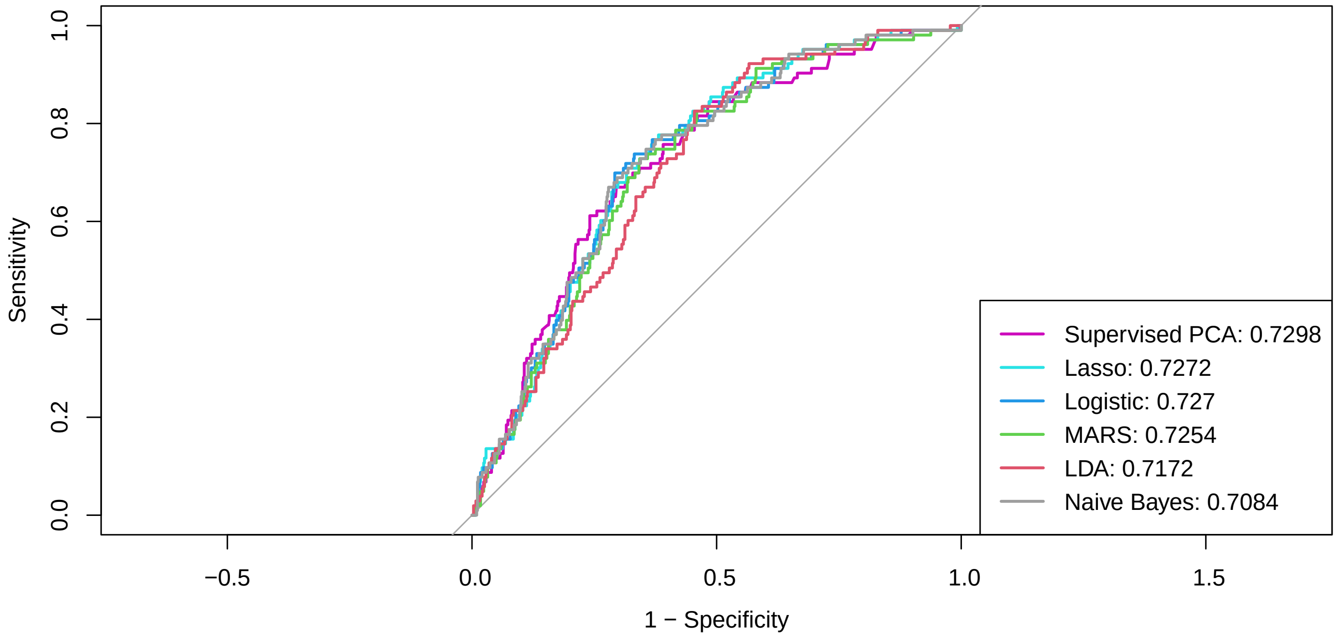
Fig.5 Distribution of the Estimated AUC for six models from 10−fold CV

Middle box line represents mean of AUC



Mean: (MARS, 0.7350), (PCA, 0.7330), (Lasso, 0.7326), (Logistic, 0.7312), (LDA, 0.7268), (NB, 0.7206)

Fig.6 AUC−ROC Curve Performance on the Test Set


Sensitivity
1 − Specificity
Supervised PCA: 0.7298
Lasso: 0.7272
Logistic: 0.727
MARS: 0.7254
LDA: 0.7172
Naive Bayes: 0.7084