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

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)ˆ2
heartRate: Heart rate (beats/minute)
glucose: Blood glucose level (mg/dL)
TenYearCHD (outcome variable): The 10 year risk of coronary heart disease(CHD).

Exploratory analysis/Parametric Model Assumption Check

From Fig.2 and Fig.3, we check whether each class has equal variance/covariance. Fig.2 shows if there exists variation in the predictors among two classes and it is to be observed the variance is unequal for **age**, **sys_bp**, **dia_bp**, **bmi**, and **glucose** variables. Fig.3 shows the covariance ellipses to check equal covariance among classes and it is to be observed the covariance ellipse for **CHD** group is generally much wider than **NoCHD** group. From Fig.4, the density of continuous predictors given outcome, $f(X|Y = y)$, are fairly normally distributed for **dia_bp**, **bmi**, **heart_rate**, and **tot_chol** variables, but not to be normally distributed for **age**, **cigs_per_day**, **glucose**, and **sys_bp** variables. It is also to be observed binary outcomes are not well dispersed by continuous predictors except **sys_bp** and **age** variables. Checking of Fig.5 for the assumptions of a logistic regression lets us see 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 analysis, we expect parametric methods, such as Logistic regression, linear discriminant analysis (LDA), or parametric (Gaussian) Naive Bayes to be not 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 our data set does not have influential outliers and there are 15 predictors (high dimensionality) among which are continuous and not highly correlated (Fig.1). 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. Table.1 shows the proportion of categorical variables by their levels of categories among response (CHD) classes. It is to be observed that variables, such as **sex**, **education**, and **prevalent_hyp**, have a fair amount of difference of proportions among classes. Lastly, note that missing data is omitted in this project in order to compare the model performance in two scenarios, omitting missing data vs. imputing missing data, which will be done in the final project.

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 in 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. For the parametric models. there are required assumptions. For logistic regression, homogeneity of variance, the normality of residuals, no influential outliers, no multicollinearity are required. For LDA, 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 is normally distributed (Fig.9) and the independence assumptions between predictors. Lastly, for Lasso, there should be no influential observations. Whether these assumptions are satisfied is aforementioned in the "Exploratory analysis/Parametric Model Assumption Check" part.

Unfortunately, most of the assumptions of the parametric models are not met with the data set; hence, this project includes two 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. AUC-ROC (ROC for caret) is the metric used to compare the performance of the models.

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), the measure of the ability of a classifier to distinguish between classes (separabiltiy), is obtained from 10-fold cross validation (repeated cv did not give better result and is more time-consuming). 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.7). 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. These five predictors are plotted for a Variance Importance plot (Fig.8) and 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 a significant role in prediction in the previous part. Lastly given that the product degree in the MARS model is 1, there is no interaction in hinge functions.

Table.2 and Fig.10 show 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 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.11). Considering that many assumptions of parametric models are violated with the data set, these results make sense. It is also not surprising the Naive Bayes method chooses to use nonparametric method with higher estimated mean value of AUC than Gaussian method (Fig.6).

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

Based on the result, 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. On the one hand, there is little discernible difference in the estimated mean of AUC between MARS and supervised PCA; hence, supervised PCA model can be also considered as the final model.

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, they can 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. For example, the plot fitted by the supervised PCA model shows that principal component variables have two different classes of response variable observations more dispersed by having 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 (they are overlapped) (Fig.8). Considering that key assumptions of LDA method are violated, this result is not surprising and hence it is to be concluded that parametric methods in machine learning are constrained to the specified functional forms and parametric assumptions. Among parametric models, Lasso regression performed the best. This makese sense reasonably considering Lasso regression is not constrained to much parametric assumptions like normality or equal covariance matrices. Thus, it is recommended to use non-parametric methods or Lasso regression method in machine learning prediction when data does not meet the assumptions, such as normality, equal variance, and independence.

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 also needs to be said that it is likely that there could be other important variables in chronic heart disease prediction that this data set does not have; hence more prediction analyses need be to done with more number of potential variables and larger number of data size 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

Table 1: Summary stat by CHD status

Characteristic	**N**	**NoCHD**, N = 2,474[1]	**CHD**, N = 454[1]	**p-value**[2]
sex	2,928			<0.001
male		43% (1,058 / 2,474)	55% (251 / 454)	
female		57% (1,416 / 2,474)	45% (203 / 454)	
age	2,928	49 (8)	54 (8)	<0.001
education	2,928			<0.001
1		39% (972 / 2,474)	53% (239 / 454)	
2		31% (770 / 2,474)	23% (103 / 454)	
3		17% (428 / 2,474)	13% (61 / 454)	
4		12% (304 / 2,474)	11% (51 / 454)	
current_smoker	2,928			0.2
0		51% (1,270 / 2,474)	48% (218 / 454)	
1		49% (1,204 / 2,474)	52% (236 / 454)	
cigs_per_day	2,928	9 (11)	11 (13)	0.016
bp_meds	2,928			<0.001
0		98% (2,414 / 2,474)	94% (425 / 454)	
1		2.4% (60 / 2,474)	6.4% (29 / 454)	
prevalent_stroke	2,928			0.002
0		100% (2,465 / 2,474)	98% (446 / 454)	
1		0.4% (9 / 2,474)	1.8% (8 / 454)	
prevalent_hyp	2,928			<0.001
0		73% (1,794 / 2,474)	48% (220 / 454)	
1		27% (680 / 2,474)	52% (234 / 454)	
diabetes	2,928			<0.001
0		98% (2,430 / 2,474)	93% (422 / 454)	
1		1.8% (44 / 2,474)	7.0% (32 / 454)	
tot_chol	2,928	235 (43)	245 (49)	<0.001
sys_bp	2,928	130 (20)	144 (27)	<0.001
dia_bp	2,928	82 (11)	87 (14)	<0.001
bmi	2,928	25.6 (3.9)	26.6 (4.6)	<0.001
heart_rate	2,928	76 (12)	77 (12)	0.057
glucose	2,928	80 (17)	90 (44)	<0.001

[1]% (n / N); Mean (SD)
[2]Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

Fig.1 Correlation plot

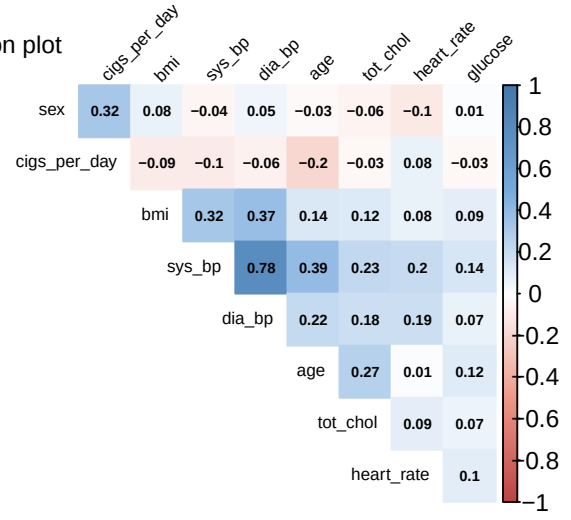

Fig.2 Checking the Assumption of Equal Variance

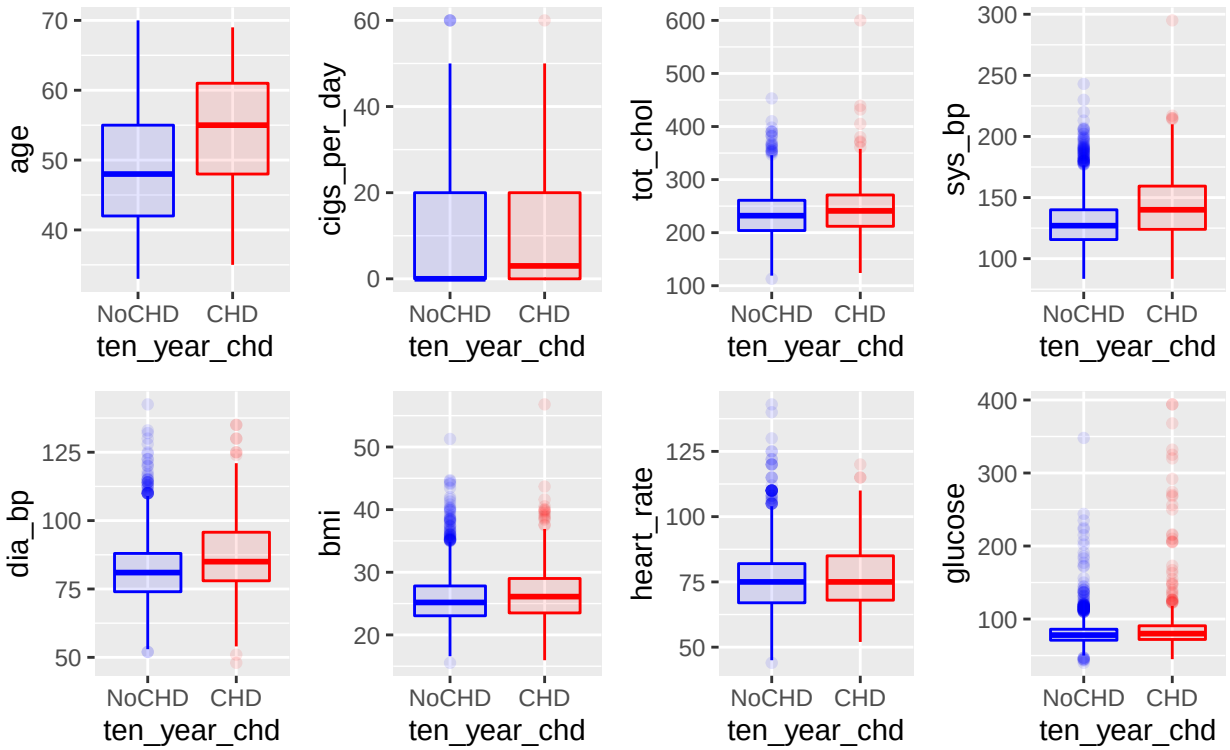

Fig.3 Checking the Assumption of Equal Covariance Ellipse

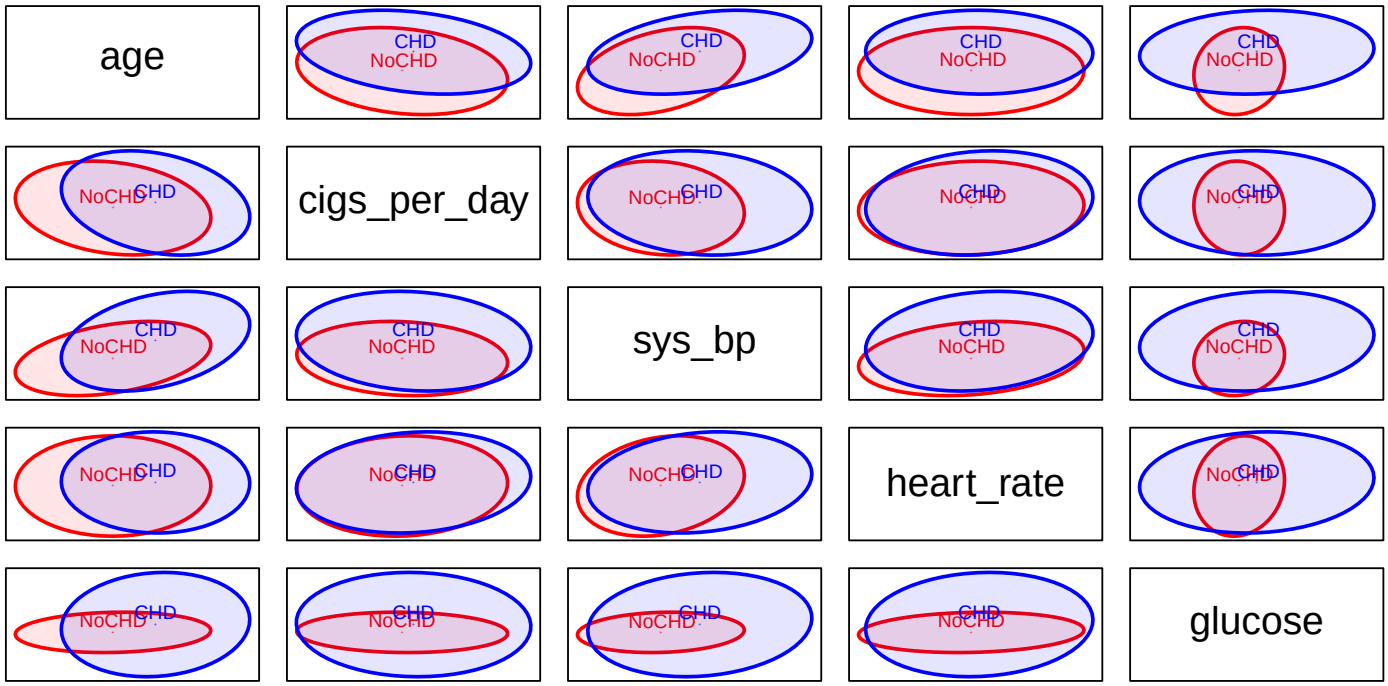

Fig.4 Checking the Assumption of Conditional Normality $f(X|Y=y) \sim N$

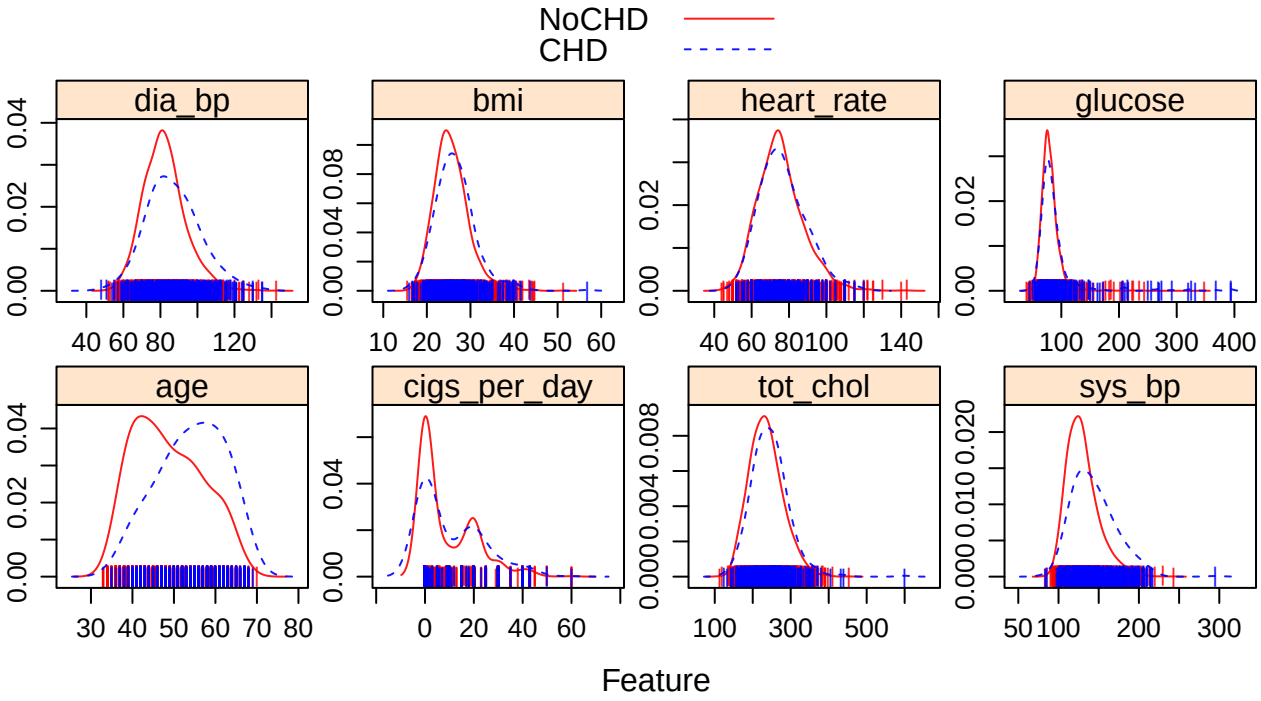

Fig.5 Checking Logistic Model Assumptions

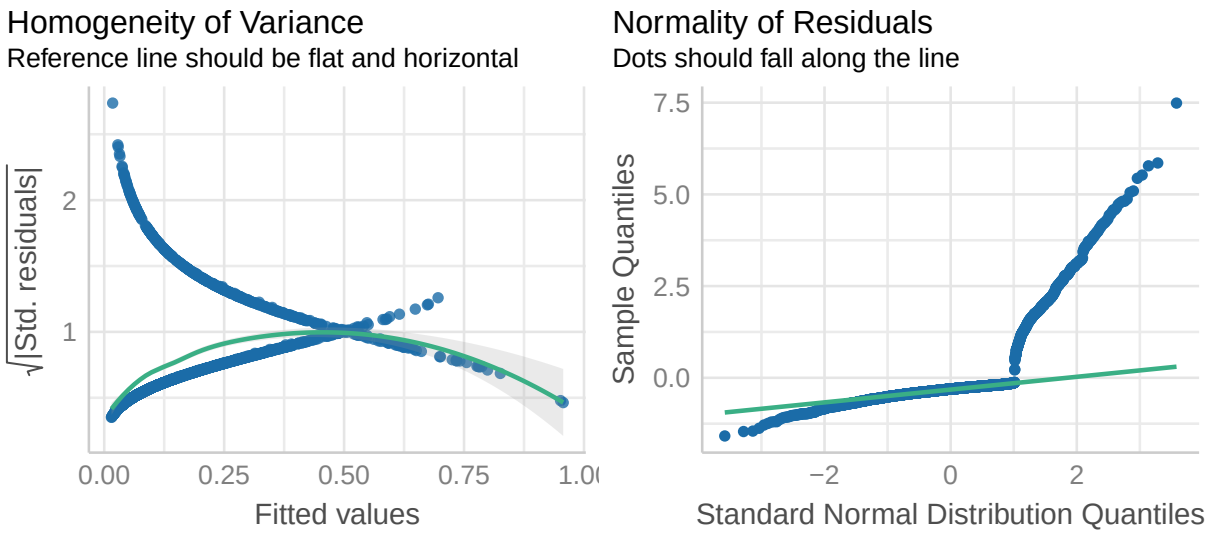

Fig.6 Model Tuning Parameter

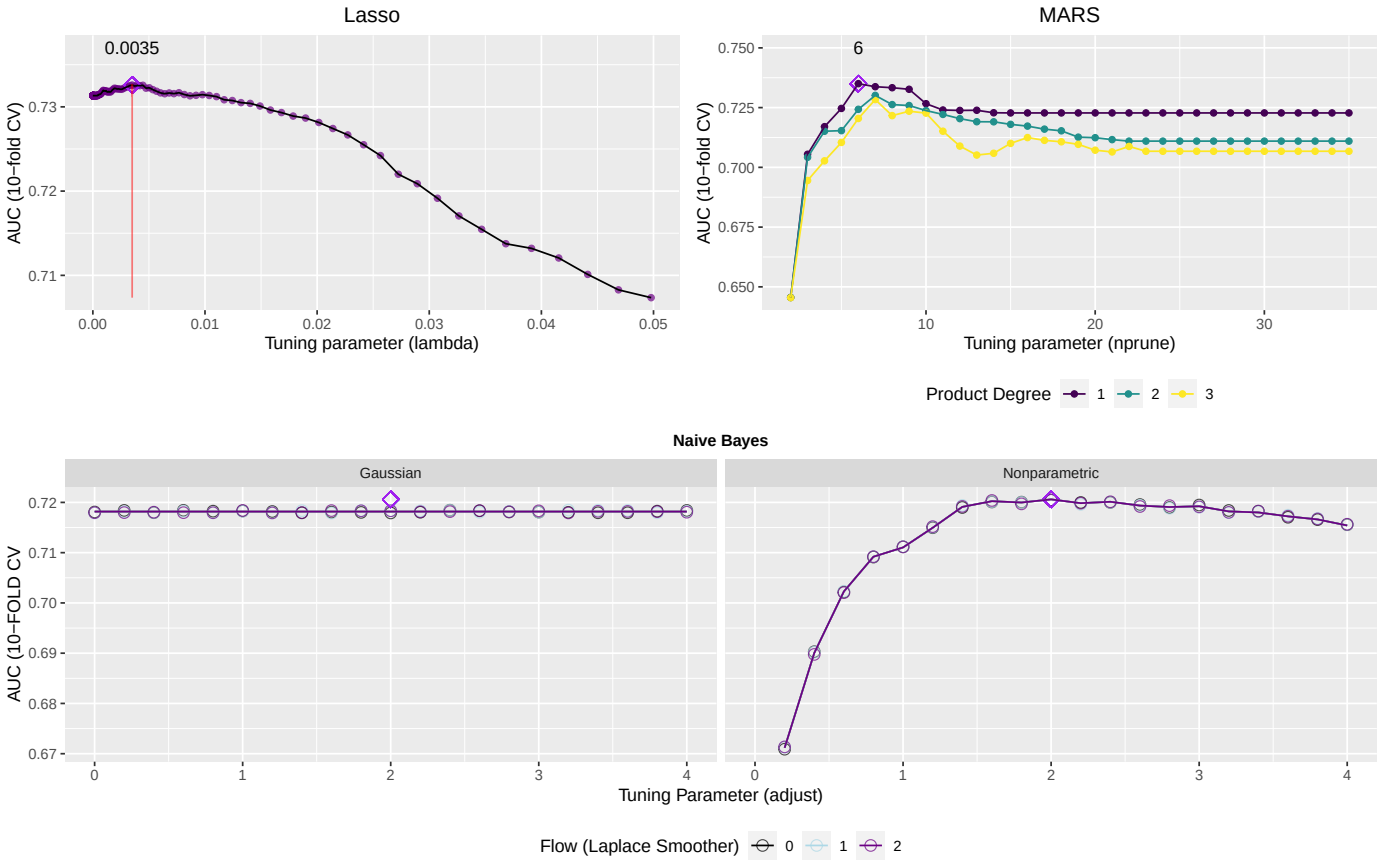

Fig.7 MARS Model Variance Importance Plot(VIP)

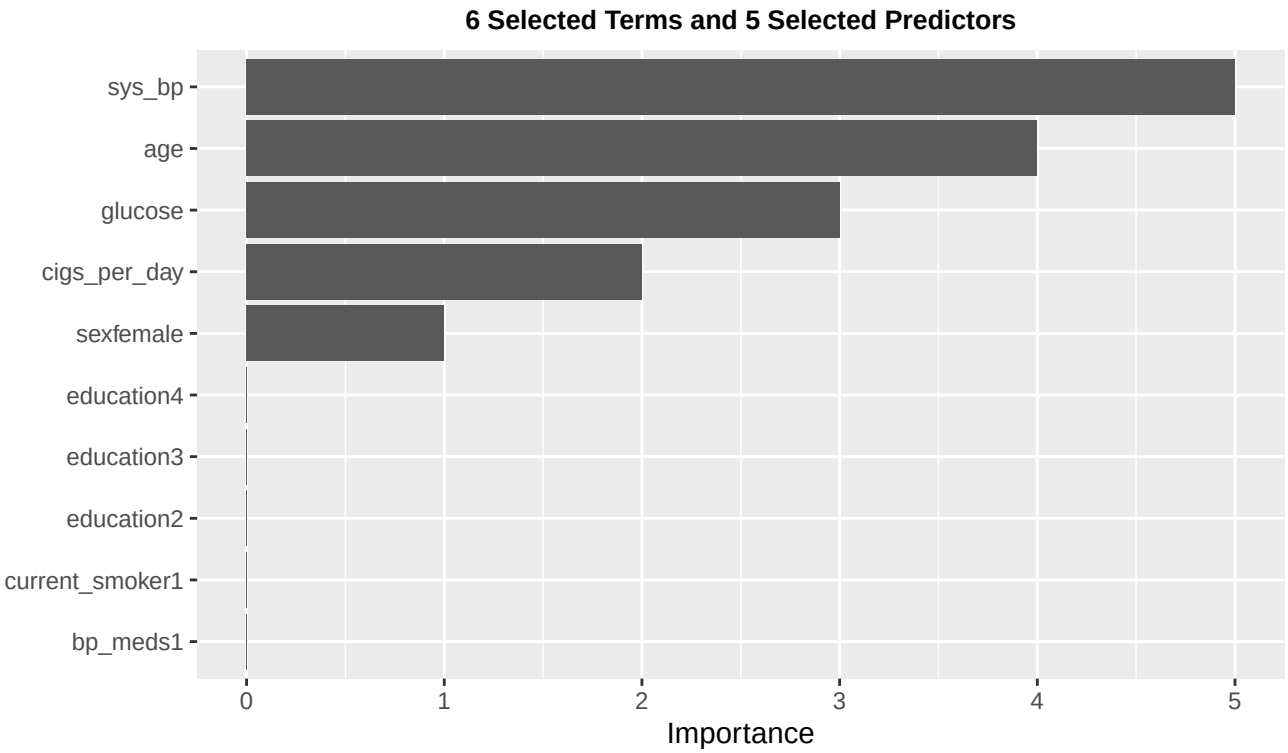

Fig.8 LDA Plot vs. PCA Plot

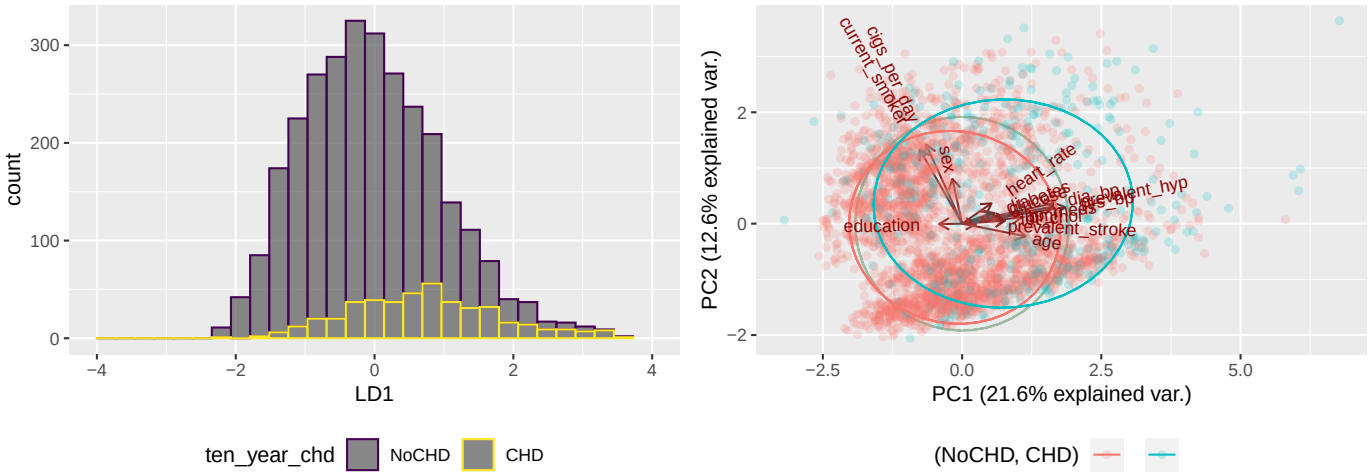

Fig.9 Naive Bayes Density Plots

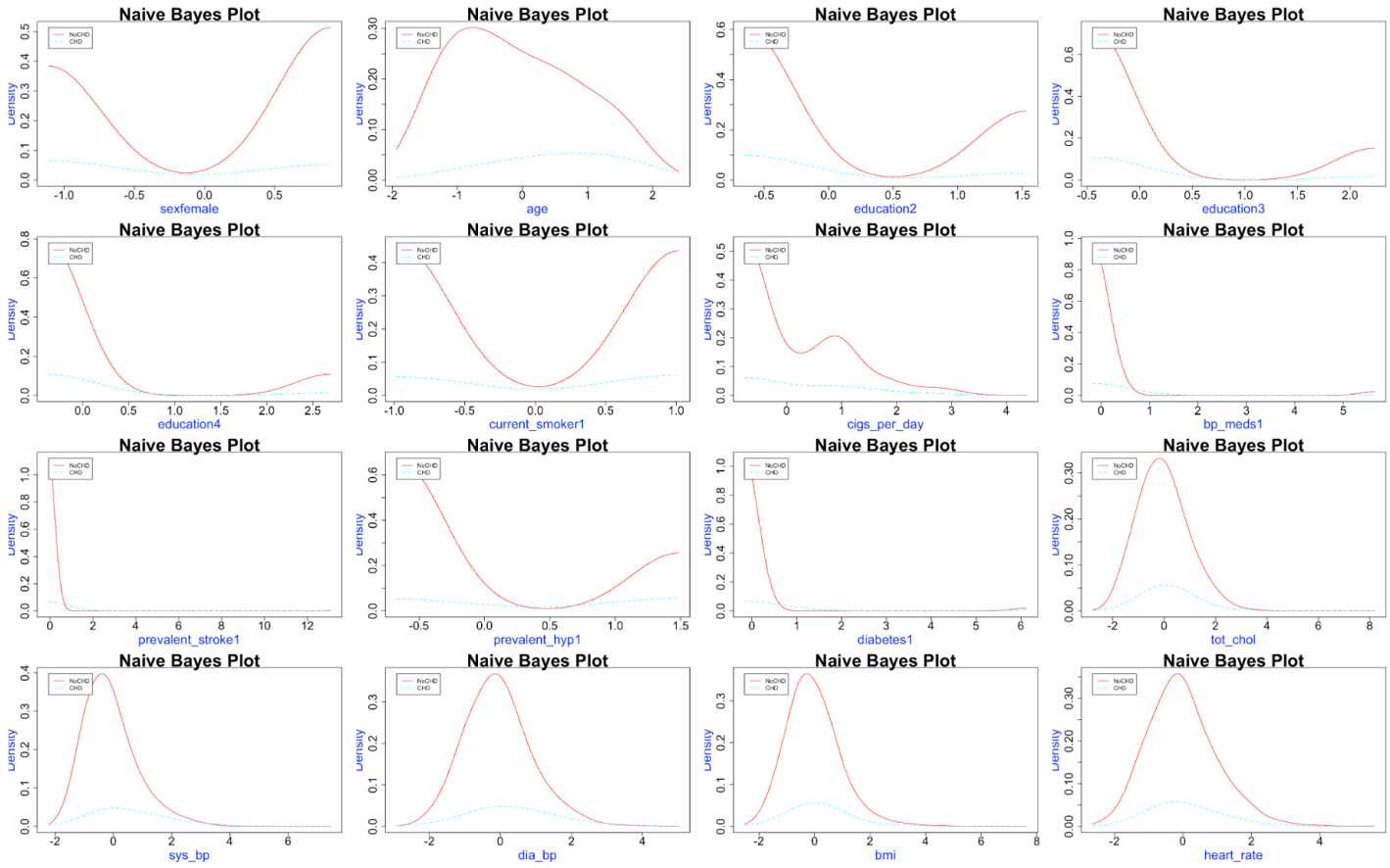

Table.2 Cross−Validated AUC Summary

Models	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
MARS	0.6710	0.7026	0.7452	0.7350	0.7554	0.7927
Supervised_pca	0.6531	0.7078	0.7392	0.7330	0.7574	0.8067
Lasso	0.6589	0.6986	0.7340	0.7326	0.7594	0.8150
Logistic	0.6639	0.6987	0.7274	0.7312	0.7613	0.8175
LDA	0.6489	0.6978	0.7268	0.7277	0.7615	0.8165
Naive_Bayes	0.6399	0.6759	0.7248	0.7206	0.7519	0.8147

Fig.10 Distribution of AUC across six models from 10−fold CV

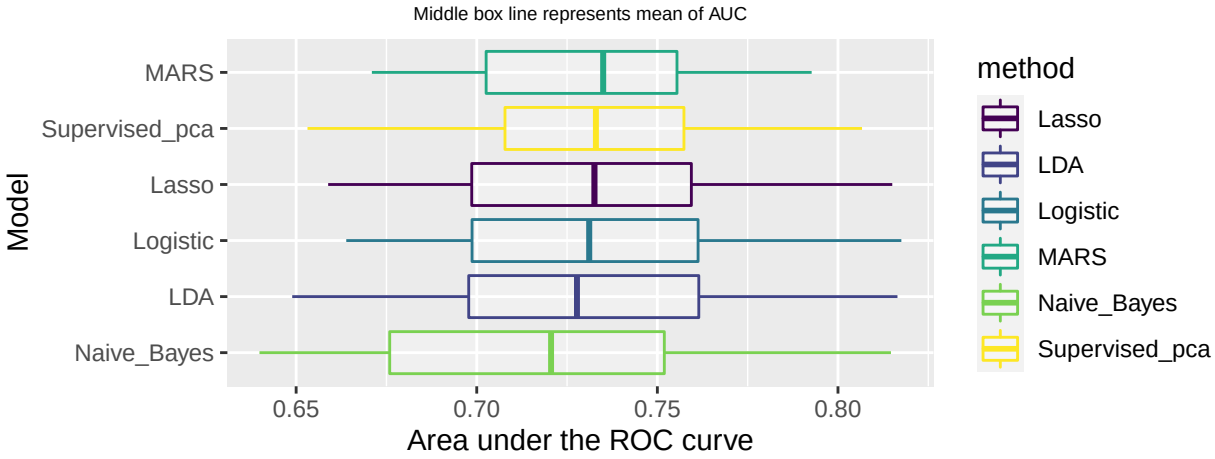

Fig.11 AUC−ROC Curve Performance on the Test Set