Stats Project 2

Theodoro Gasperin Terra Camargo

2024-05-13

### Question 1

Study and describe the predictor variables. Do you see any issues that are relevant for making predictions? Make sure to discuss the dimensionality of the data and the implication on fitting models.

n > p 5 misssing data dimensionality of the data corinality <- variables that are correlated need to be removed

## Loading required package: lattice

### Question 2

Fit and compare an appropriate unconstrained linear model, as well as lasso and ridge regression models. Discuss what you find. What is an appropriate base-level of performance to compare your models to?

## Loading required package: Matrix

## Loaded glmnet 4.1-8

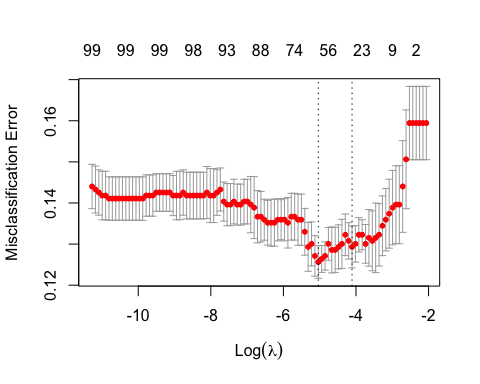
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## unconstrained\_pred\_class  
## FALSE TRUE  
## FALSE 252 33  
## TRUE 29 25

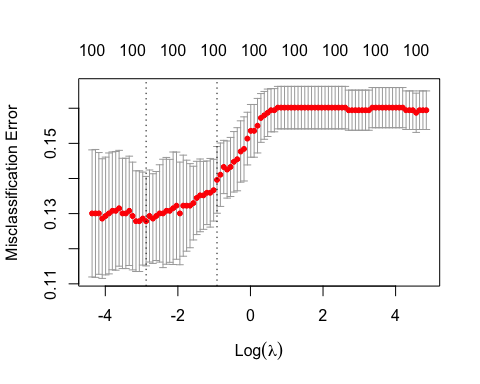


## [1] 0.006482767

## [1] 0.01643618

## 101 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -3.366556772  
## AGE 0.032059346  
## SEX -0.032815476  
## INF\_ANAM .   
## STENOK\_AN 0.089034500  
## FK\_STENOK .   
## IBS\_POST 0.199270080  
## IBS\_NASL -0.394719011  
## GB 0.138265067  
## SIM\_GIPERT 0.478794427  
## DLIT\_AG 0.006441915  
## ZSN\_A 0.356549350  
## nr11 -0.548841204  
## nr01 -1.630902959  
## nr02 .   
## nr03 0.072477604  
## nr04 0.335730946  
## nr07 .   
## nr08 -0.391516253  
## np01 0.295236284  
## np04 .   
## np05 -1.120475985  
## np07 .   
## np08 0.360309111  
## np09 .   
## np10 0.170404476  
## endocr\_01 0.071806459  
## endocr\_02 1.037643185  
## endocr\_03 -1.123013956  
## zab\_leg\_01 -0.381694794  
## zab\_leg\_02 0.416531812  
## zab\_leg\_03 0.386249230  
## zab\_leg\_04 .   
## zab\_leg\_06 .   
## S\_AD\_ORIT -0.002619525  
## D\_AD\_ORIT -0.008193716  
## O\_L\_POST 0.164194017  
## K\_SH\_POST 1.973945666  
## MP\_TP\_POST .   
## SVT\_POST 0.927031354  
## GT\_POST -0.513611985  
## FIB\_G\_POST .   
## ant\_im 0.390949535  
## lat\_im 0.120493256  
## inf\_im 0.253163325  
## post\_im 0.222547077  
## IM\_PG\_P 1.142413502  
## ritm\_ecg\_p\_01 -0.608149075  
## ritm\_ecg\_p\_02 .   
## ritm\_ecg\_p\_04 0.001489239  
## ritm\_ecg\_p\_06 -3.304073199  
## ritm\_ecg\_p\_07 0.012776893  
## ritm\_ecg\_p\_08 -0.328375806  
## n\_r\_ecg\_p\_01 0.248512325  
## n\_r\_ecg\_p\_02 .   
## n\_r\_ecg\_p\_03 .   
## n\_r\_ecg\_p\_04 0.495136728  
## n\_r\_ecg\_p\_05 .   
## n\_r\_ecg\_p\_06 .   
## n\_r\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_09 .   
## n\_r\_ecg\_p\_10 -1.170687183  
## n\_p\_ecg\_p\_01 -0.137708836  
## n\_p\_ecg\_p\_03 0.552910876  
## n\_p\_ecg\_p\_04 2.078890947  
## n\_p\_ecg\_p\_05 .   
## n\_p\_ecg\_p\_06 .   
## n\_p\_ecg\_p\_07 .   
## n\_p\_ecg\_p\_08 1.245035579  
## n\_p\_ecg\_p\_09 .   
## n\_p\_ecg\_p\_10 0.399437062  
## n\_p\_ecg\_p\_11 .   
## n\_p\_ecg\_p\_12 0.869176628  
## fibr\_ter\_01 0.002172043  
## fibr\_ter\_02 .   
## fibr\_ter\_03 0.227233281  
## fibr\_ter\_05 .   
## fibr\_ter\_06 .   
## fibr\_ter\_07 .   
## fibr\_ter\_08 0.421095017  
## GIPO\_K .   
## K\_BLOOD .   
## GIPER\_Na .   
## Na\_BLOOD .   
## ALT\_BLOOD .   
## AST\_BLOOD .   
## KFK\_BLOOD -0.013769765  
## L\_BLOOD 0.052008825  
## ROE 0.008492154  
## TIME\_B\_S -0.118464541  
## NA\_KB -0.526347793  
## NOT\_NA\_KB -0.317362904  
## LID\_KB -0.258325951  
## NITR\_S 1.019824300  
## LID\_S\_n .   
## B\_BLOK\_S\_n -0.297439112  
## ANT\_CA\_S\_n -0.253870511  
## GEPAR\_S\_n .   
## ASP\_S\_n -0.723350806  
## TIKL\_S\_n .   
## TRENT\_S\_n .

## 101 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -2.795318e+00  
## AGE 2.484686e-02  
## SEX .   
## INF\_ANAM .   
## STENOK\_AN 6.002922e-02  
## FK\_STENOK .   
## IBS\_POST 9.957241e-02  
## IBS\_NASL -6.633706e-01  
## GB .   
## SIM\_GIPERT .   
## DLIT\_AG 1.943041e-02  
## ZSN\_A 3.051517e-01  
## nr11 .   
## nr01 .   
## nr02 .   
## nr03 .   
## nr04 7.026289e-02  
## nr07 .   
## nr08 .   
## np01 .   
## np04 .   
## np05 .   
## np07 .   
## np08 .   
## np09 .   
## np10 .   
## endocr\_01 .   
## endocr\_02 5.120634e-01  
## endocr\_03 .   
## zab\_leg\_01 -1.057433e-02  
## zab\_leg\_02 2.192196e-01  
## zab\_leg\_03 .   
## zab\_leg\_04 .   
## zab\_leg\_06 .   
## S\_AD\_ORIT -7.820391e-05  
## D\_AD\_ORIT -7.801039e-03  
## O\_L\_POST 1.782457e-01  
## K\_SH\_POST 1.686431e+00  
## MP\_TP\_POST .   
## SVT\_POST .   
## GT\_POST .   
## FIB\_G\_POST .   
## ant\_im 2.306077e-01  
## lat\_im 1.450265e-02  
## inf\_im 1.069103e-01  
## post\_im 2.504616e-02  
## IM\_PG\_P 7.977742e-01  
## ritm\_ecg\_p\_01 -4.744407e-01  
## ritm\_ecg\_p\_02 .   
## ritm\_ecg\_p\_04 .   
## ritm\_ecg\_p\_06 -6.108712e-01  
## ritm\_ecg\_p\_07 .   
## ritm\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_01 .   
## n\_r\_ecg\_p\_02 .   
## n\_r\_ecg\_p\_03 .   
## n\_r\_ecg\_p\_04 2.394166e-02  
## n\_r\_ecg\_p\_05 .   
## n\_r\_ecg\_p\_06 .   
## n\_r\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_09 .   
## n\_r\_ecg\_p\_10 .   
## n\_p\_ecg\_p\_01 .   
## n\_p\_ecg\_p\_03 2.254582e-01  
## n\_p\_ecg\_p\_04 .   
## n\_p\_ecg\_p\_05 .   
## n\_p\_ecg\_p\_06 .   
## n\_p\_ecg\_p\_07 .   
## n\_p\_ecg\_p\_08 .   
## n\_p\_ecg\_p\_09 .   
## n\_p\_ecg\_p\_10 3.869474e-02  
## n\_p\_ecg\_p\_11 .   
## n\_p\_ecg\_p\_12 5.773620e-01  
## fibr\_ter\_01 .   
## fibr\_ter\_02 .   
## fibr\_ter\_03 .   
## fibr\_ter\_05 .   
## fibr\_ter\_06 .   
## fibr\_ter\_07 .   
## fibr\_ter\_08 .   
## GIPO\_K .   
## K\_BLOOD .   
## GIPER\_Na .   
## Na\_BLOOD .   
## ALT\_BLOOD .   
## AST\_BLOOD .   
## KFK\_BLOOD .   
## L\_BLOOD 3.475240e-02  
## ROE 2.542876e-03  
## TIME\_B\_S -6.656695e-02  
## NA\_KB -7.793126e-02  
## NOT\_NA\_KB -1.085085e-01  
## LID\_KB -6.699495e-02  
## NITR\_S 7.821847e-01  
## LID\_S\_n .   
## B\_BLOK\_S\_n .   
## ANT\_CA\_S\_n -9.267728e-02  
## GEPAR\_S\_n .   
## ASP\_S\_n -5.360968e-01  
## TIKL\_S\_n .   
## TRENT\_S\_n .



## [1] 0.05638381

## [1] 0.3977766

## 101 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -3.366556772  
## AGE 0.032059346  
## SEX -0.032815476  
## INF\_ANAM .   
## STENOK\_AN 0.089034500  
## FK\_STENOK .   
## IBS\_POST 0.199270080  
## IBS\_NASL -0.394719011  
## GB 0.138265067  
## SIM\_GIPERT 0.478794427  
## DLIT\_AG 0.006441915  
## ZSN\_A 0.356549350  
## nr11 -0.548841204  
## nr01 -1.630902959  
## nr02 .   
## nr03 0.072477604  
## nr04 0.335730946  
## nr07 .   
## nr08 -0.391516253  
## np01 0.295236284  
## np04 .   
## np05 -1.120475985  
## np07 .   
## np08 0.360309111  
## np09 .   
## np10 0.170404476  
## endocr\_01 0.071806459  
## endocr\_02 1.037643185  
## endocr\_03 -1.123013956  
## zab\_leg\_01 -0.381694794  
## zab\_leg\_02 0.416531812  
## zab\_leg\_03 0.386249230  
## zab\_leg\_04 .   
## zab\_leg\_06 .   
## S\_AD\_ORIT -0.002619525  
## D\_AD\_ORIT -0.008193716  
## O\_L\_POST 0.164194017  
## K\_SH\_POST 1.973945666  
## MP\_TP\_POST .   
## SVT\_POST 0.927031354  
## GT\_POST -0.513611985  
## FIB\_G\_POST .   
## ant\_im 0.390949535  
## lat\_im 0.120493256  
## inf\_im 0.253163325  
## post\_im 0.222547077  
## IM\_PG\_P 1.142413502  
## ritm\_ecg\_p\_01 -0.608149075  
## ritm\_ecg\_p\_02 .   
## ritm\_ecg\_p\_04 0.001489239  
## ritm\_ecg\_p\_06 -3.304073199  
## ritm\_ecg\_p\_07 0.012776893  
## ritm\_ecg\_p\_08 -0.328375806  
## n\_r\_ecg\_p\_01 0.248512325  
## n\_r\_ecg\_p\_02 .   
## n\_r\_ecg\_p\_03 .   
## n\_r\_ecg\_p\_04 0.495136728  
## n\_r\_ecg\_p\_05 .   
## n\_r\_ecg\_p\_06 .   
## n\_r\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_09 .   
## n\_r\_ecg\_p\_10 -1.170687183  
## n\_p\_ecg\_p\_01 -0.137708836  
## n\_p\_ecg\_p\_03 0.552910876  
## n\_p\_ecg\_p\_04 2.078890947  
## n\_p\_ecg\_p\_05 .   
## n\_p\_ecg\_p\_06 .   
## n\_p\_ecg\_p\_07 .   
## n\_p\_ecg\_p\_08 1.245035579  
## n\_p\_ecg\_p\_09 .   
## n\_p\_ecg\_p\_10 0.399437062  
## n\_p\_ecg\_p\_11 .   
## n\_p\_ecg\_p\_12 0.869176628  
## fibr\_ter\_01 0.002172043  
## fibr\_ter\_02 .   
## fibr\_ter\_03 0.227233281  
## fibr\_ter\_05 .   
## fibr\_ter\_06 .   
## fibr\_ter\_07 .   
## fibr\_ter\_08 0.421095017  
## GIPO\_K .   
## K\_BLOOD .   
## GIPER\_Na .   
## Na\_BLOOD .   
## ALT\_BLOOD .   
## AST\_BLOOD .   
## KFK\_BLOOD -0.013769765  
## L\_BLOOD 0.052008825  
## ROE 0.008492154  
## TIME\_B\_S -0.118464541  
## NA\_KB -0.526347793  
## NOT\_NA\_KB -0.317362904  
## LID\_KB -0.258325951  
## NITR\_S 1.019824300  
## LID\_S\_n .   
## B\_BLOK\_S\_n -0.297439112  
## ANT\_CA\_S\_n -0.253870511  
## GEPAR\_S\_n .   
## ASP\_S\_n -0.723350806  
## TIKL\_S\_n .   
## TRENT\_S\_n .

## 101 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -2.795318e+00  
## AGE 2.484686e-02  
## SEX .   
## INF\_ANAM .   
## STENOK\_AN 6.002922e-02  
## FK\_STENOK .   
## IBS\_POST 9.957241e-02  
## IBS\_NASL -6.633706e-01  
## GB .   
## SIM\_GIPERT .   
## DLIT\_AG 1.943041e-02  
## ZSN\_A 3.051517e-01  
## nr11 .   
## nr01 .   
## nr02 .   
## nr03 .   
## nr04 7.026289e-02  
## nr07 .   
## nr08 .   
## np01 .   
## np04 .   
## np05 .   
## np07 .   
## np08 .   
## np09 .   
## np10 .   
## endocr\_01 .   
## endocr\_02 5.120634e-01  
## endocr\_03 .   
## zab\_leg\_01 -1.057433e-02  
## zab\_leg\_02 2.192196e-01  
## zab\_leg\_03 .   
## zab\_leg\_04 .   
## zab\_leg\_06 .   
## S\_AD\_ORIT -7.820391e-05  
## D\_AD\_ORIT -7.801039e-03  
## O\_L\_POST 1.782457e-01  
## K\_SH\_POST 1.686431e+00  
## MP\_TP\_POST .   
## SVT\_POST .   
## GT\_POST .   
## FIB\_G\_POST .   
## ant\_im 2.306077e-01  
## lat\_im 1.450265e-02  
## inf\_im 1.069103e-01  
## post\_im 2.504616e-02  
## IM\_PG\_P 7.977742e-01  
## ritm\_ecg\_p\_01 -4.744407e-01  
## ritm\_ecg\_p\_02 .   
## ritm\_ecg\_p\_04 .   
## ritm\_ecg\_p\_06 -6.108712e-01  
## ritm\_ecg\_p\_07 .   
## ritm\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_01 .   
## n\_r\_ecg\_p\_02 .   
## n\_r\_ecg\_p\_03 .   
## n\_r\_ecg\_p\_04 2.394166e-02  
## n\_r\_ecg\_p\_05 .   
## n\_r\_ecg\_p\_06 .   
## n\_r\_ecg\_p\_08 .   
## n\_r\_ecg\_p\_09 .   
## n\_r\_ecg\_p\_10 .   
## n\_p\_ecg\_p\_01 .   
## n\_p\_ecg\_p\_03 2.254582e-01  
## n\_p\_ecg\_p\_04 .   
## n\_p\_ecg\_p\_05 .   
## n\_p\_ecg\_p\_06 .   
## n\_p\_ecg\_p\_07 .   
## n\_p\_ecg\_p\_08 .   
## n\_p\_ecg\_p\_09 .   
## n\_p\_ecg\_p\_10 3.869474e-02  
## n\_p\_ecg\_p\_11 .   
## n\_p\_ecg\_p\_12 5.773620e-01  
## fibr\_ter\_01 .   
## fibr\_ter\_02 .   
## fibr\_ter\_03 .   
## fibr\_ter\_05 .   
## fibr\_ter\_06 .   
## fibr\_ter\_07 .   
## fibr\_ter\_08 .   
## GIPO\_K .   
## K\_BLOOD .   
## GIPER\_Na .   
## Na\_BLOOD .   
## ALT\_BLOOD .   
## AST\_BLOOD .   
## KFK\_BLOOD .   
## L\_BLOOD 3.475240e-02  
## ROE 2.542876e-03  
## TIME\_B\_S -6.656695e-02  
## NA\_KB -7.793126e-02  
## NOT\_NA\_KB -1.085085e-01  
## LID\_KB -6.699495e-02  
## NITR\_S 7.821847e-01  
## LID\_S\_n .   
## B\_BLOK\_S\_n .   
## ANT\_CA\_S\_n -9.267728e-02  
## GEPAR\_S\_n .   
## ASP\_S\_n -5.360968e-01  
## TIKL\_S\_n .   
## TRENT\_S\_n .

## Confusion Matrix and Statistics  
##   
## unconstrained\_pred\_class  
## FALSE TRUE  
## FALSE 252 33  
## TRUE 29 25  
##   
## Accuracy : 0.8171   
## 95% CI : (0.7718, 0.8568)  
## No Information Rate : 0.8289   
## P-Value [Acc > NIR] : 0.7448   
##   
## Kappa : 0.3371   
##   
## Mcnemar's Test P-Value : 0.7032   
##   
## Sensitivity : 0.8968   
## Specificity : 0.4310   
## Pos Pred Value : 0.8842   
## Neg Pred Value : 0.4630   
## Prevalence : 0.8289   
## Detection Rate : 0.7434   
## Detection Prevalence : 0.8407   
## Balanced Accuracy : 0.6639   
##   
## 'Positive' Class : FALSE   
##

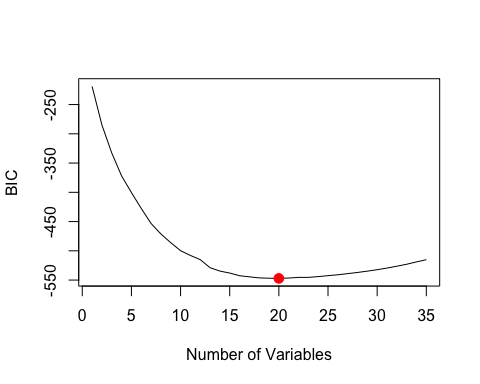
## Confusion Matrix and Statistics  
##   
## lasso\_pred.min  
## FALSE TRUE  
## FALSE 276 9  
## TRUE 30 24  
##   
## Accuracy : 0.885   
## 95% CI : (0.8461, 0.9169)  
## No Information Rate : 0.9027   
## P-Value [Acc > NIR] : 0.881493   
##   
## Kappa : 0.4901   
##   
## Mcnemar's Test P-Value : 0.001362   
##   
## Sensitivity : 0.9020   
## Specificity : 0.7273   
## Pos Pred Value : 0.9684   
## Neg Pred Value : 0.4444   
## Prevalence : 0.9027   
## Detection Rate : 0.8142   
## Detection Prevalence : 0.8407   
## Balanced Accuracy : 0.8146   
##   
## 'Positive' Class : FALSE   
##

## Confusion Matrix and Statistics  
##   
## ridge\_pred.min  
## FALSE TRUE  
## FALSE 277 8  
## TRUE 33 21  
##   
## Accuracy : 0.8791   
## 95% CI : (0.8395, 0.9118)  
## No Information Rate : 0.9145   
## P-Value [Acc > NIR] : 0.9897731   
##   
## Kappa : 0.4442   
##   
## Mcnemar's Test P-Value : 0.0001781   
##   
## Sensitivity : 0.8935   
## Specificity : 0.7241   
## Pos Pred Value : 0.9719   
## Neg Pred Value : 0.3889   
## Prevalence : 0.9145   
## Detection Rate : 0.8171   
## Detection Prevalence : 0.8407   
## Balanced Accuracy : 0.8088   
##   
## 'Positive' Class : FALSE   
##

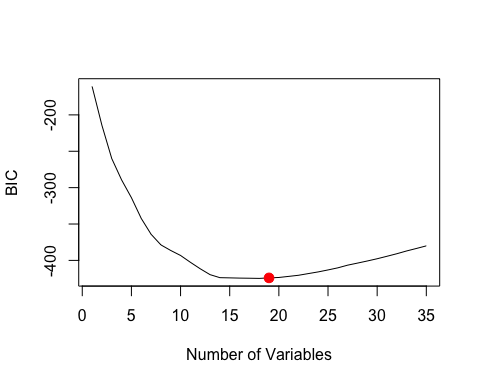
### Question 3

Among your top predictors, do you see evidence of non-linear effects? How could you accommodate non-linear effects and still use a regularized regression approach? Does adding non-linear effects improve your model?

## [1] 20



## [1] 18



Fit a non-linear regression (e.g. spline model like GAM) and then compare it to the linear model using AIC or likelihood ratio test. This is a simple and intuitive method of testing non-linearity. If the test rejects, or if AIC prefers the GAM, then conclude there are non-linearities.

## Loading required package: carData

##   
## Attaching package: 'car'

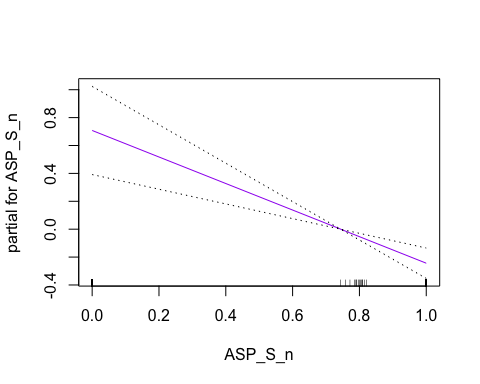
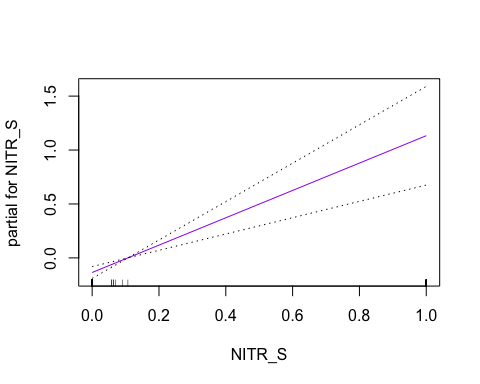
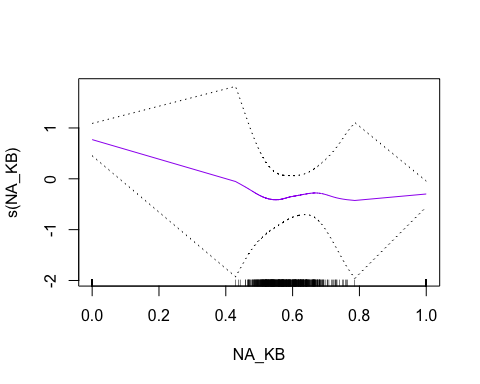
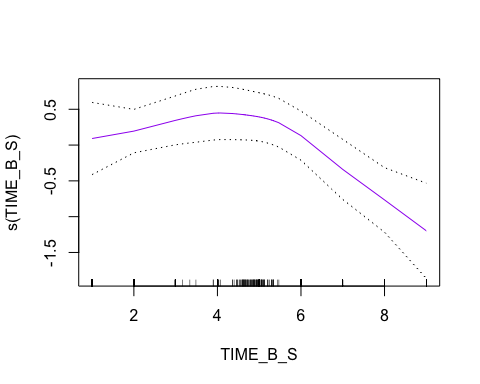
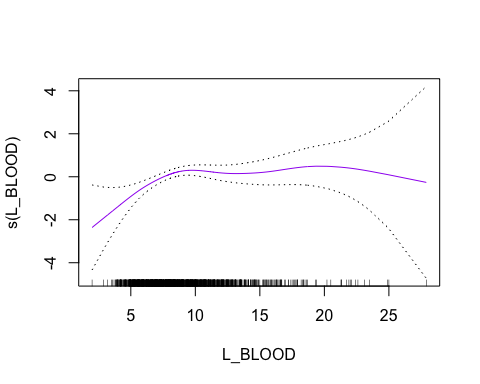
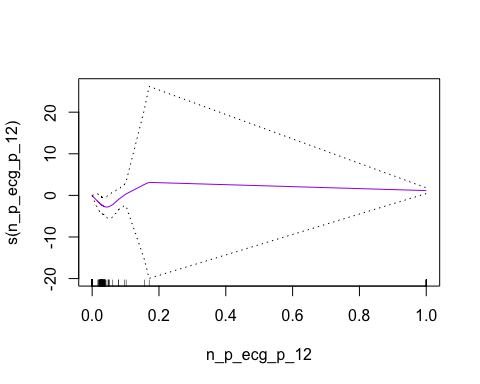
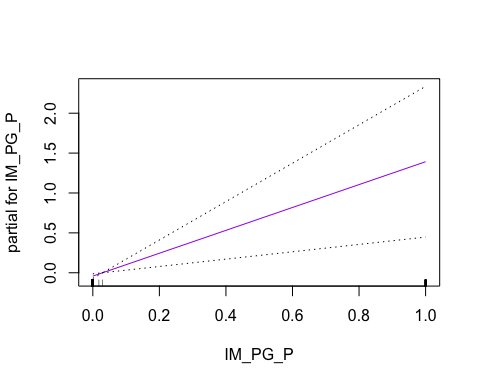
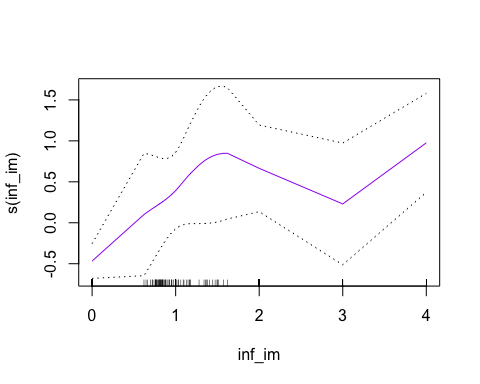
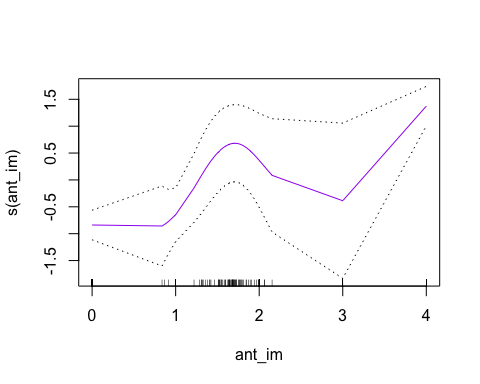
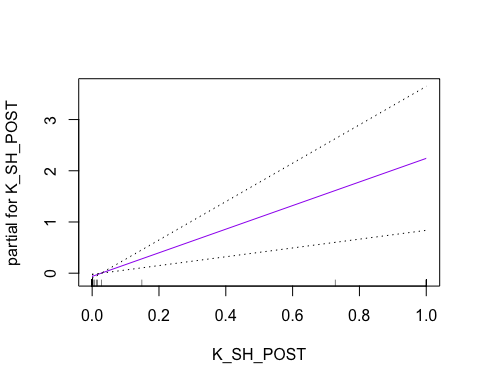
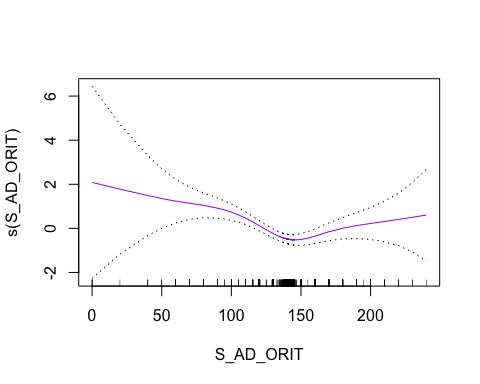
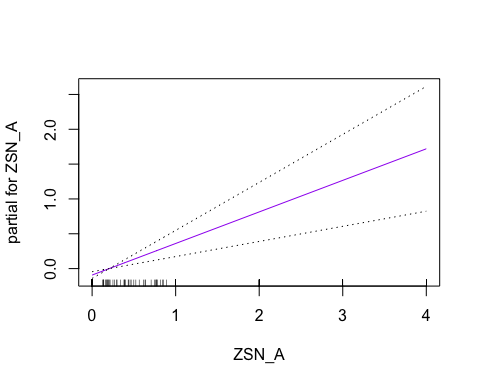
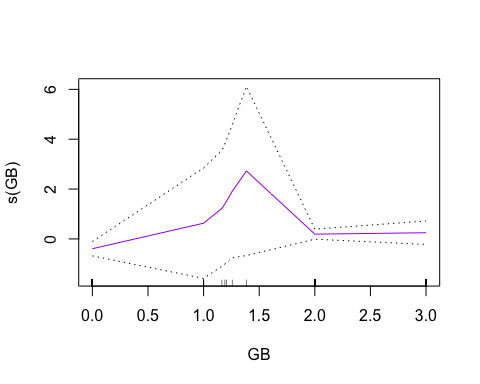
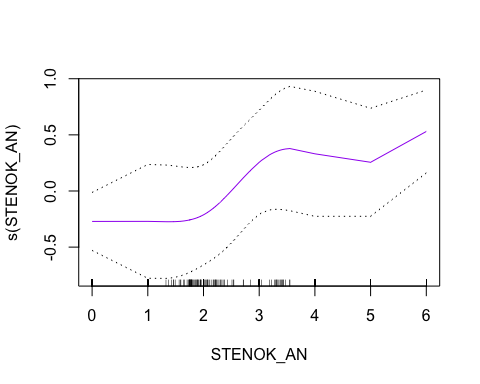
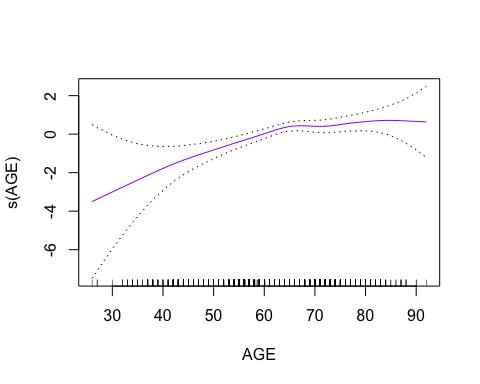
## The following object is masked from 'package:dplyr':  
##   
## recode

## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.22-3

##   
## Call: gam(formula = LET\_IS ~ s(AGE) + s(STENOK\_AN) + s(GB) + ZSN\_A +   
## s(S\_AD\_ORIT) + K\_SH\_POST + s(ant\_im) + s(inf\_im) + IM\_PG\_P +   
## s(n\_p\_ecg\_p\_12) + s(L\_BLOOD) + s(TIME\_B\_S) + s(NA\_KB) + NITR\_S +   
## ASP\_S\_n, family = binomial, data = MI\_train)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -1.93174 -0.41347 -0.22953 -0.08526 2.92365   
##   
## (Dispersion Parameter for binomial family taken to be 1)  
##   
## Null Deviance: 1194.258 on 1360 degrees of freedom  
## Residual Deviance: 678.3789 on 1315 degrees of freedom  
## AIC: 770.3792   
##   
## Number of Local Scoring Iterations: NA   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## s(AGE) 1 18.82 18.8247 24.4005 8.830e-07 \*\*\*  
## s(STENOK\_AN) 1 5.49 5.4922 7.1190 0.0077210 \*\*   
## s(GB) 1 1.16 1.1595 1.5029 0.2204476   
## ZSN\_A 1 9.07 9.0674 11.7531 0.0006262 \*\*\*  
## s(S\_AD\_ORIT) 1 16.06 16.0562 20.8120 5.539e-06 \*\*\*  
## K\_SH\_POST 1 8.91 8.9096 11.5486 0.0006983 \*\*\*  
## s(ant\_im) 1 28.84 28.8354 37.3763 1.282e-09 \*\*\*  
## s(inf\_im) 1 14.45 14.4495 18.7294 1.620e-05 \*\*\*  
## IM\_PG\_P 1 7.35 7.3513 9.5287 0.0020651 \*\*   
## s(n\_p\_ecg\_p\_12) 1 9.82 9.8181 12.7263 0.0003735 \*\*\*  
## s(L\_BLOOD) 1 4.68 4.6813 6.0679 0.0138932 \*   
## s(TIME\_B\_S) 1 4.80 4.8034 6.2262 0.0127091 \*   
## s(NA\_KB) 1 12.45 12.4485 16.1357 6.230e-05 \*\*\*  
## NITR\_S 1 22.93 22.9251 29.7154 5.969e-08 \*\*\*  
## ASP\_S\_n 1 20.07 20.0730 26.0185 3.877e-07 \*\*\*  
## Residuals 1315 1014.51 0.7715   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Anova for Nonparametric Effects  
## Npar Df Npar Chisq P(Chi)   
## (Intercept)   
## s(AGE) 3 8.2884 0.040411 \*   
## s(STENOK\_AN) 3 2.6212 0.453795   
## s(GB) 3 4.1120 0.249625   
## ZSN\_A   
## s(S\_AD\_ORIT) 3 22.3431 5.532e-05 \*\*\*  
## K\_SH\_POST   
## s(ant\_im) 3 9.6805 0.021487 \*   
## s(inf\_im) 3 6.2603 0.099616 .   
## IM\_PG\_P   
## s(n\_p\_ecg\_p\_12) 3 7.2024 0.065718 .   
## s(L\_BLOOD) 3 11.1444 0.010972 \*   
## s(TIME\_B\_S) 3 11.4846 0.009374 \*\*   
## s(NA\_KB) 3 5.2840 0.152159   
## NITR\_S   
## ASP\_S\_n   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



## Confusion Matrix and Statistics  
##   
## gam\_pred  
## FALSE TRUE  
## FALSE 276 9  
## TRUE 29 25  
##   
## Accuracy : 0.8879   
## 95% CI : (0.8494, 0.9194)  
## No Information Rate : 0.8997   
## P-Value [Acc > NIR] : 0.794478   
##   
## Kappa : 0.5076   
##   
## Mcnemar's Test P-Value : 0.002055   
##   
## Sensitivity : 0.9049   
## Specificity : 0.7353   
## Pos Pred Value : 0.9684   
## Neg Pred Value : 0.4630   
## Prevalence : 0.8997   
## Detection Rate : 0.8142   
## Detection Prevalence : 0.8407   
## Balanced Accuracy : 0.8201   
##   
## 'Positive' Class : FALSE   
##

### Question 4

Fit an appropriate Random Forest model. Report a comparison of performance to your linear model and explain any differences in performance. Do you see an important difference in how variables are used for predictions?

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

## Confusion Matrix and Statistics  
##   
## forest\_pred  
## FALSE TRUE  
## FALSE 284 1  
## TRUE 35 19  
##   
## Accuracy : 0.8938   
## 95% CI : (0.856, 0.9245)  
## No Information Rate : 0.941   
## P-Value [Acc > NIR] : 0.9997   
##   
## Kappa : 0.4677   
##   
## Mcnemar's Test P-Value : 3.798e-08   
##   
## Sensitivity : 0.8903   
## Specificity : 0.9500   
## Pos Pred Value : 0.9965   
## Neg Pred Value : 0.3519   
## Prevalence : 0.9410   
## Detection Rate : 0.8378   
## Detection Prevalence : 0.8407   
## Balanced Accuracy : 0.9201   
##   
## 'Positive' Class : FALSE   
##

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases

## Setting levels: control = FALSE, case = TRUE

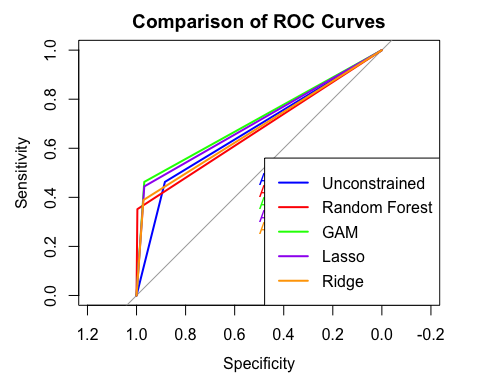
## Setting direction: controls < cases

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases

 The ROC curve comparison reveals significant differences in the performance of the models based on the Area Under the Curve (AUC) metric. The Generalized Additive Model (GAM) demonstrates the highest AUC of 0.716, indicating superior discriminative ability compared to the other models. The Lasso regression follows closely with an AUC of 0.706, suggesting that it also performs well in differentiating between classes. The Ridge regression, with an AUC of 0.680, performs moderately, surpassing the Unconstrained linear model and the Random Forest, both of which have an AUC of 0.674. This indicates that while regularization techniques like Lasso and Ridge improve model performance over the basic linear model, the GAM’s incorporation of non-linear effects provides the most substantial enhancement in predictive capability. The Random Forest model’s performance being on par with the Unconstrained model suggests that it may not be capturing complex relationships in the data as effectively as expected. Overall, these findings highlight the importance of considering non-linear effects and regularization techniques to enhance model accuracy and reliability in predictive tasks.