Assignment1

Tiffany Woodley

5/25/2020

Question 3: GAMs and MARS

The data utilised for this question comes from a dataset which measures the quality of red wines. The dataset however contains no information about the grapes, wine brand, selling price ect and only contains physiocochemical predictor variables with a sensory response variable. The data will be used for the purpose of classification between good and bad wines.

The physiocochemical predictor variables are measured on different scales which will be taken into account when trying to model interactions.

The ratings of the wines range between 3 and 8, it was taken that a wine was of good quality if it was above the average rating (5.636) or bad quality if it was below average.

Splitting based on a wine being above or below average resulted in a pretty even class distribution with 53.47% being above average and 46.52% being below.

This question will therefore explore Generalised Additive Models(GAMs) and Multivariate Adaptive Regression Splines(MARS) for the purpose of differentiating between good and bad wines.

Generalised Additive Models

GAMs are fit using the function **gam** from the **mgcv** library. Since the response is binary the errors are assumed to be from a binomial distribution and the logit link function is used. The **gam** function's built in **s** function is used to fit smooth splines to the predictor variables with **bs = cr** indicating that cubic regression splines should be used. The built in **te** function was used to model interactions between predictor variables also with **bs=cr** indicating that cubic regression splines should be used for the marginal basis functions. Penalisation terms are left to be calculated automatically using the function's default **method = GCVcp** which uses GCV for known scales and Mallow's Cp for unknown scales in order to choose the best fitting penalisation values(λ).

Since GAMs do not allow for automatic variable selection, four models are fitted to assess which variables/interaction would be useful. The first model was fitted using smooths of all 11 of the predictor variables ignoring interactions between them.

Looking at the summary of the model it is clear that there are some smooths that do not hold high significance in the model, and so are just making the model more complex unecessarily.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ s(fixed.acidity) + s(volatile.acidity) + s(citric.acid) +
##     s(residual.sugar) + s(chlorides) + s(free.sulfur.dioxide) +
##     s(total.sulfur.dioxide) + s(density) + s(pH) + s(sulphates) +
##     s(alcohol)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

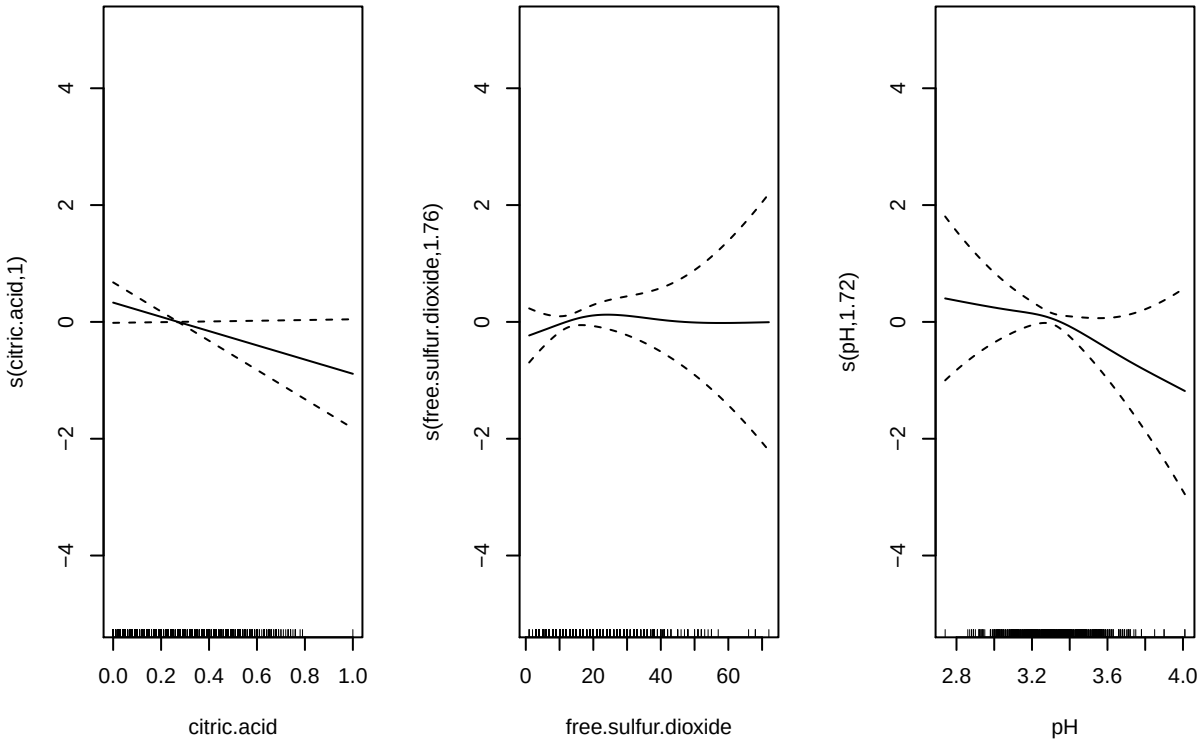
```
## (Intercept)   0.4085     0.1490   2.743   0.0061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(fixed.acidity)        8.806  8.979 20.743   0.0163 *
## s(volatile.acidity)     1.005  1.009 24.290 8.83e-07 ***
## s(citric.acid)          1.001  1.002  3.635   0.0566 .
## s(residual.sugar)       8.337  8.833 14.998   0.0838 .
## s(chlorides)            7.347  8.071 20.092   0.0104 *
## s(free.sulfur.dioxide)  1.755  2.239  1.580   0.5606
## s(total.sulfur.dioxide) 6.874  7.236 36.312 8.71e-06 ***
## s(density)              8.625  8.883 15.673   0.0678 .
## s(pH)                   1.716  2.211  3.262   0.2376
## s(sulphates)            3.909  4.864 91.159  < 2e-16 ***
## s(alcohol)              6.360  7.268 64.677 2.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.388   Deviance explained = 34.9%
## UBRE = -0.029409  Scale est. = 1         n = 1599
```

To test whether we can improve the model, we remove variables one at a time that have the least significance until all smooths within the model are likely significant. The resultant model (second model) contained 8 smooths and the summary can be seen below.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ s(fixed.acidity) + s(volatile.acidity) + s(residual.sugar) +
##     s(chlorides) + s(total.sulfur.dioxide) + s(density) + s(sulphates) +
##     s(alcohol)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4077     0.1497   2.723  0.00647 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(fixed.acidity)        4.485  5.560  19.42  0.00308 **
## s(volatile.acidity)     1.000  1.000  26.37 2.82e-07 ***
## s(residual.sugar)       8.501  8.900  15.75  0.07020 .
## s(chlorides)            6.407  7.545  19.45  0.01031 *
## s(total.sulfur.dioxide) 6.946  7.300  39.75 1.76e-06 ***
## s(density)              8.689  8.915  19.46  0.02069 *
## s(sulphates)            4.101  5.077  94.24  < 2e-16 ***
## s(alcohol)              6.349  7.399  63.46 7.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## R-sq.(adj) =  0.376   Deviance explained = 33.8%
## UBRE = -0.026401  Scale est. = 1         n = 1599
```

Below we can see the three smooths that the were removed from the model, the smooths do not vary from zero much and so would not bring a lot of information into the model.



The third model contains all potential smooths and interaction terms (using tensot product splines as the variables are on different scales). Since it would be computationally infeasible to add all the possible interaction terms into the model, each potential interaction term is added to the model one at a time and the p value is looked at to see if it suggests significance. Although it would be more accurate to add in all the terms, this was done for the sake of computational efficiency.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ s(fixed.acidity) + s(volatile.acidity) + s(citric.acid) +
##     s(residual.sugar) + s(chlorides) + s(free.sulfur.dioxide) +
##     s(total.sulfur.dioxide) + s(density) + s(pH) + s(sulphates) +
##     s(alcohol) + te(fixed.acidity, citric.acid) + te(fixed.acidity,
##     chlorides) + te(fixed.acidity, sulphates) + te(fixed.acidity,
##     alcohol) + te(volatile.acidity, sulphates) + te(citric.acid,
##     residual.sugar) + te(citric.acid, alcohol) + te(residual.sugar,
##     chlorides) + te(chlorides, total.sulfur.dioxide) + te(chlorides,
##     density) + te(chlorides, sulphates) + te(chlorides, alcohol) +
##     te(free.sulfur.dioxide, density) + te(free.sulfur.dioxide,
##     sulphates) + te(total.sulfur.dioxide, sulphates) + te(pH,
##     sulphates) + te(sulphates, alcohol)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)   0.5957     0.2055   2.899  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                       edf Ref.df Chi.sq  p-value
## s(fixed.acidity)                 1.000e+00  1.000  3.816 0.050779 .
## s(volatile.acidity)              7.676e+00  7.934 33.998 3.02e-05 ***
## s(citric.acid)                   1.000e+00  1.000  0.045 0.832606
## s(residual.sugar)                9.000e+00  9.000 11.353 0.251993
## s(chlorides)                     1.000e+00  1.000  0.014 0.904830
## s(free.sulfur.dioxide)           1.823e+00  2.283  1.855 0.510574
## s(total.sulfur.dioxide)          6.114e+00  6.608 12.320 0.077227 .
## s(density)                       8.683e+00  8.910 25.336 0.002908 **
## s(pH)                            1.000e+00  1.000  0.171 0.679232
## s(sulphates)                     1.000e+00  1.000  0.082 0.774677
## s(alcohol)                       4.150e+00  5.057 20.332 0.000792 ***
## te(fixed.acidity,citric.acid)    4.342e+00  4.932  2.086 0.865013
## te(fixed.acidity,chlorides)      7.980e+00 20.000 17.848 0.002168 **
## te(fixed.acidity,sulphates)      1.233e+01 20.000 27.681 0.000431 ***
## te(fixed.acidity,alcohol)        4.175e+00 19.000 23.181 4.05e-05 ***
## te(volatile.acidity,sulphates)   1.729e-05 20.000  0.000 1.000000
## te(citric.acid,residual.sugar)   1.222e+01 20.000 26.643 0.001902 **
## te(citric.acid,alcohol)          7.980e+00 16.000 23.542 0.000622 ***
## te(residual.sugar,chlorides)     1.952e+00 20.000  2.694 0.173588
## te(chlorides,total.sulfur.dioxide) 6.890e-01 10.000  2.120 0.066743 .
## te(chlorides,density)            1.248e-04 20.000  0.000 0.226397
## te(chlorides,sulphates)          3.759e-05 16.000  0.000 0.654288
## te(chlorides,alcohol)            5.891e+00 17.000 11.239 0.008396 **
## te(free.sulfur.dioxide,density)  5.890e-01 20.000  0.976 0.144787
## te(free.sulfur.dioxide,sulphates) 9.884e-05 18.000  0.000 0.341948
## te(total.sulfur.dioxide,sulphates) 3.290e+00 16.000  8.578 0.014320 *
## te(pH,sulphates)                 6.842e+00 19.000 15.930 0.001277 **
## te(sulphates,alcohol)            1.486e-03 17.000  0.001 0.422390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.469   Deviance explained = 44.5%
## UBRE = -0.093771  Scale est. = 1         n = 1599
```

For the fourth model, terms indicating low significance are removed from the model. This model contains 9 additive terms, the inclusion of interaction terms allows for more smooths to be removed as they capture a fuller relationship of those variables.

Deviance measures the difference between the log likelihood of the saturated (perfect fit) model and the fitted model. If deviance is equal to 0 the fitted model fits the data perfectly and so lower deviance values indicate a better fit to the training data.

For comparison between the four models we look at the deviance (the goodness of fit) and the are under the ROC curve (prediction accuracy) on unseen data. The models are fit on 75% of the data and then tested on 25% of the data.

These fitted models are then used to predict the test data, which will give us a better idea about their prediction accuracy (generalisation).

The table below compares the 4 GAM models

	Deviance	AUC.train	AUC.test	No..terms
GAM 1	1060.7446	0.8726052	0.8154091	12
GAM 2	1070.4699	0.8690137	0.8142738	8
GAM 3	947.2915	0.9010216	0.8390979	25
GAM 4	992.6873	0.8895182	0.8244658	8

GAM 3 had the best fitting model with the lowest deviance and the highest prediction accuracy. All models had big differences between the train and test auc indicating the models have overfit to the training data. The reduced term models GAM 2 and GAM 4 had extremely similar prediction accuracy as the models containing all the terms, they also had higher deviance values. This shows that the extra terms which were not shown to be significant did not bring much more information to the model, and so the models with fewer terms produce similar results with a much smaller interprable model.

Multivariate Adaptive Regression Splines

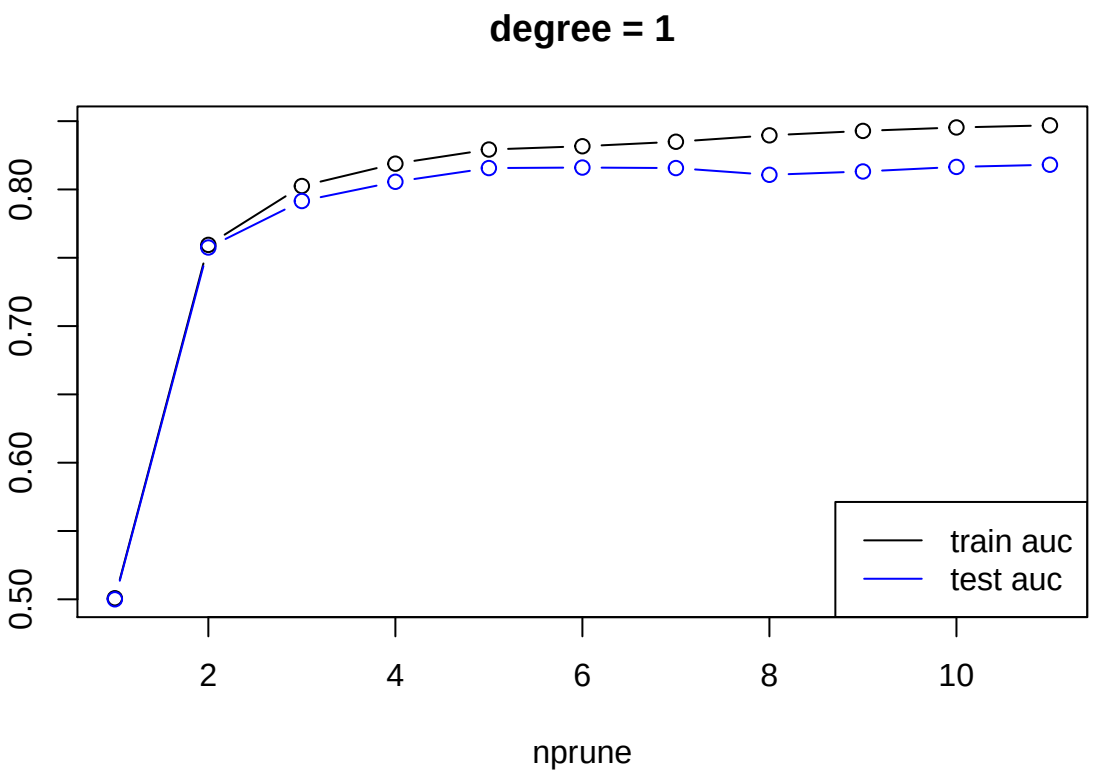
Next we fit a MARS model to the data, this model also uses a link function with the assumption of binomial errors for the pupose of classification. Multiple models are fit in order to find the model with the best trade off between fit and prediction accuracy. This model performs variable selection automatically, making the process a simpler one than using GAMs. These models are fit using the **earth** function from the **earth** library.

The first two models were fitted with **degree = 1** and **degree = 2** with the GVC penalty per knot being 2 and 3 respectively. The maximum number of model terms created by the forward pass was left as the default **nk = min(200 max(20 2ncol(x))) + 1**. Since the threshold was left at **thresh = 0.001**, the threshold will likely be reached before the maximum number of terms is hit. The pruning method was set to **method = backward** and number of pruned terms were not limited with **nprune = NULL**.

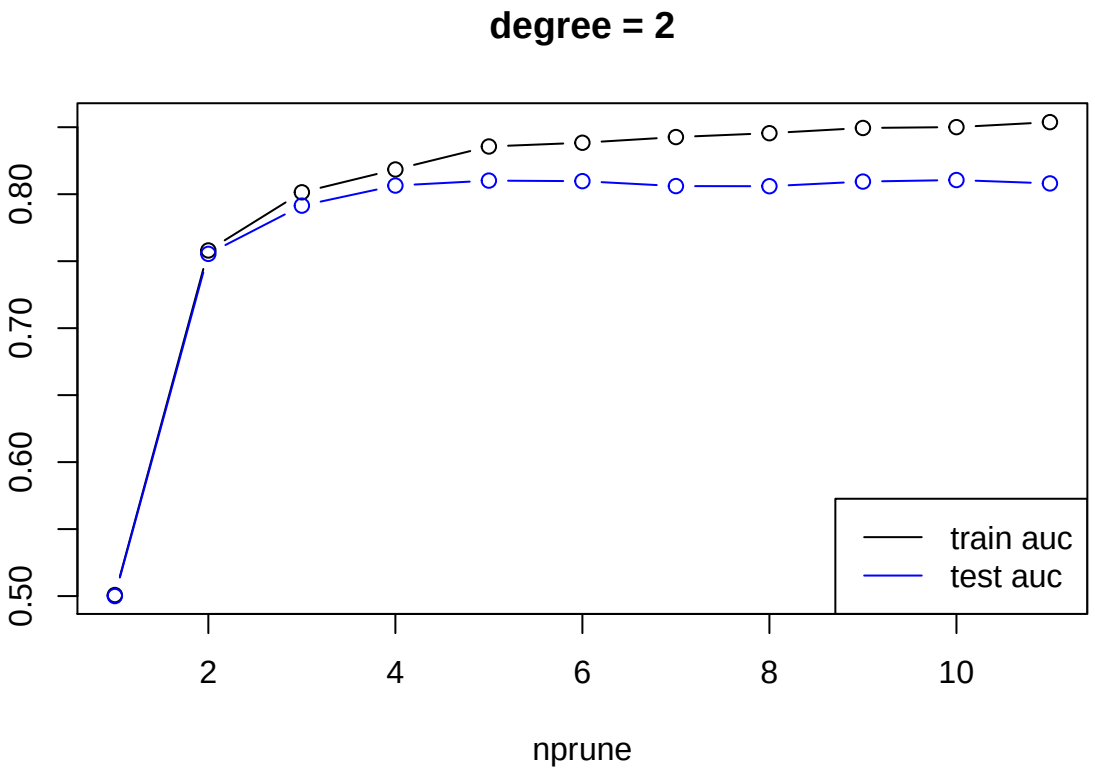
The model with degree = 2 (MARS 2) has a better fit to the training data with lower deviance and higher auc train, but perform worse in the test auc indicating overfitting. The degree = 1 model (MARS 1) has slightly higher deviance, lower training auc and slightly higher test auc indicating a better model for generalisation. There is still quite a large discrepency between the test and train aucs and so the parameter **nprune** was looked into to see if a better model could be found.

	Deviance	AUC.train	AUC.test	No.terms
MARS 1	1158.513	0.8469257	0.8181084	11
MARS 2	1131.563	0.8553948	0.8030727	12

For the degree = 1 model, at **nprune = 6** the test auc stops improving whilst the train auc increases and therefore **nprune = 5** was chosen.



For the degree = 2 model, at **nprune = 5** the test auc stops improving whilst the train auc increases and therefore **nprune = 4** was chosen.



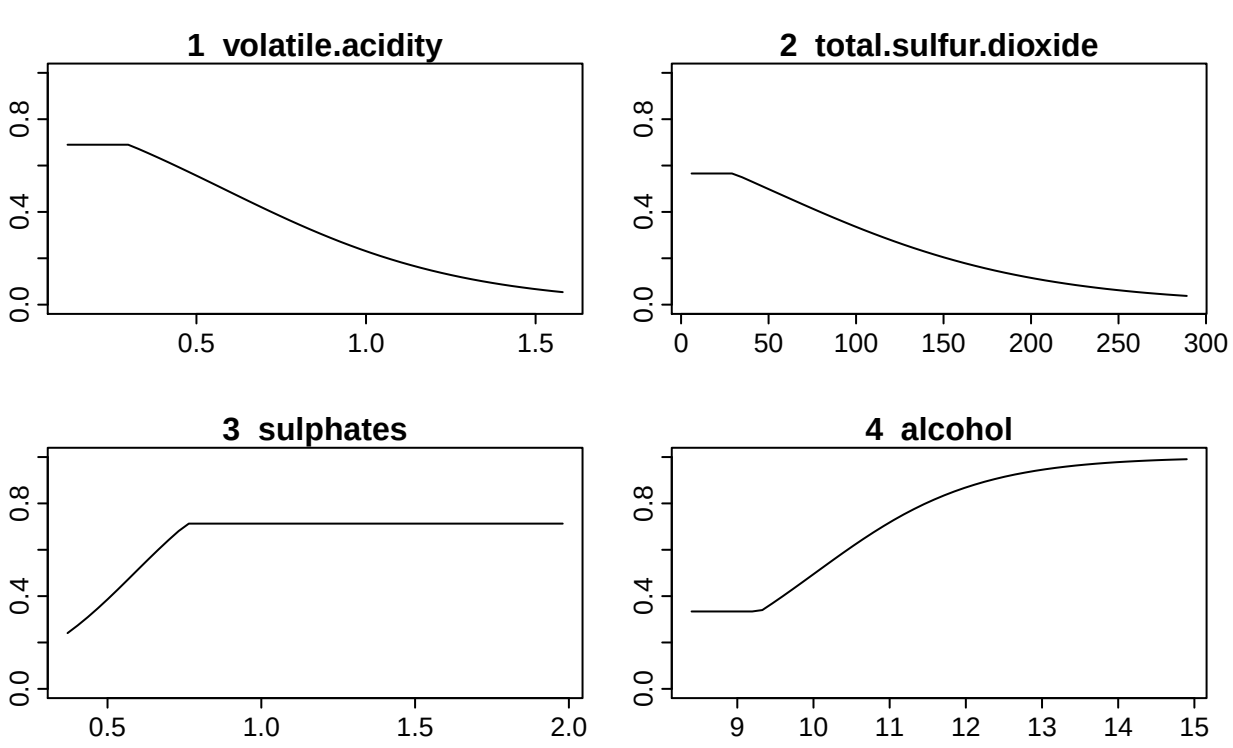
The pruned models have similar test and train aucs indicating good generalisation. The test auc values are similar to the unpruned models but the train auc has come down and the deviance has increased. This

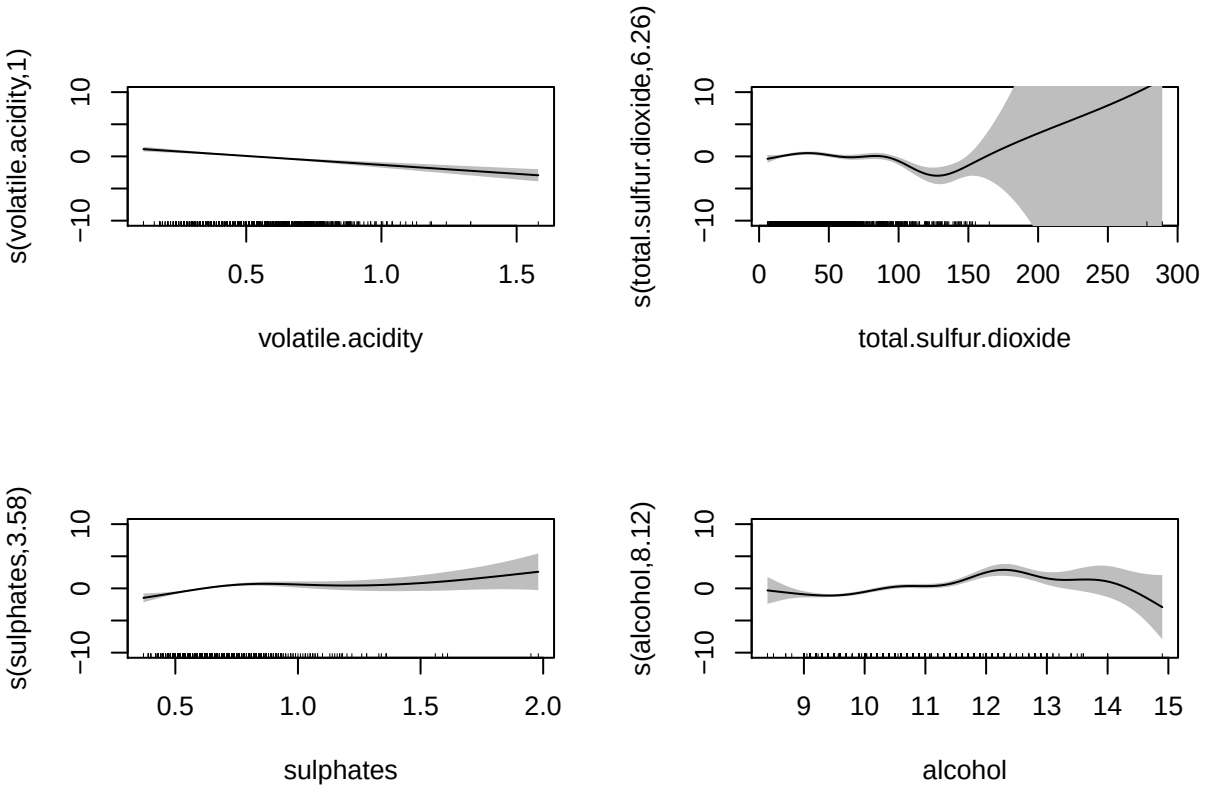
means we have not removed any relevent variables from the model. The model with degree = 1 (MARS 3) outperforms the model with degree = 2 (MARS 4) and thus was chosen as the final MARS model.

	Deviance	AUC.train	AUC.test	No..terms
MARS 3	1218.766	0.8292336	0.8156613	5
MARS 4	1247.221	0.8184395	0.8064406	4

A GAM model was fit on the same 4 variables kept in MARS 3 with an intercept term for comparison. The fitted terms can be visualised below, with the top 4 graphs belonging to the MARS 3 model and the bottom four belonging to the GAM fitted on the same variables. Both models capture similar relationships whilst the GAM's relationships are more wiggly due to it's higher flexibility.

```
##  plotmo grid:    fixed.acidity volatile.acidity citric.acid residual.sugar
##                            7.9              0.52        0.26            2.2
##  chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
##      0.079                  13                   37  0.9967 3.31      0.62
##  alcohol quality
##     10.2       6
```





	Deviance	AUC.train	AUC.test	No..of.terms
GAM Model	1156.876	0.8459259	0.8197482	5
MARS Model	1218.766	0.8292336	0.8156613	5

The GAM model, fitted on the same 4 terms as the MARs model found significant, has a lower training auc compared to previous fitted GAMs whilst still maintaining a similar test auc indicating a model with better generalisation, and so was chosen as the final GAM model. This goes to show that MARs does a good job in identifying variable importance. Between the final models the GAM has the better goodness of fit (lower deviance and higher train auc) and a slightly higher prediction accuracy (test auc). Looking at the plotted fits the GAM provides more wiggly functions explaining the higher goodness of fit. The MARs model with a lower goodness of fit (higher devaince and lower train auc) has very similar test auc values.

References

1. Erni, Birgit, 2020, *Chapter 3: Generalised Additive Models(GAMs)*, lecture notes, Advanced Analytics STA5057W, University of Cape Town, May 2020

2. Erni, Birgit, 2020, *Chapter 4: Multivariate Adaptive Regression Splines(MARS)*, lecture notes, Advanced Analytics STA5057W, University of Cape Town, May 2020