Assignment 2

Valentin Kodderitzsch 3895157

2024-05-17

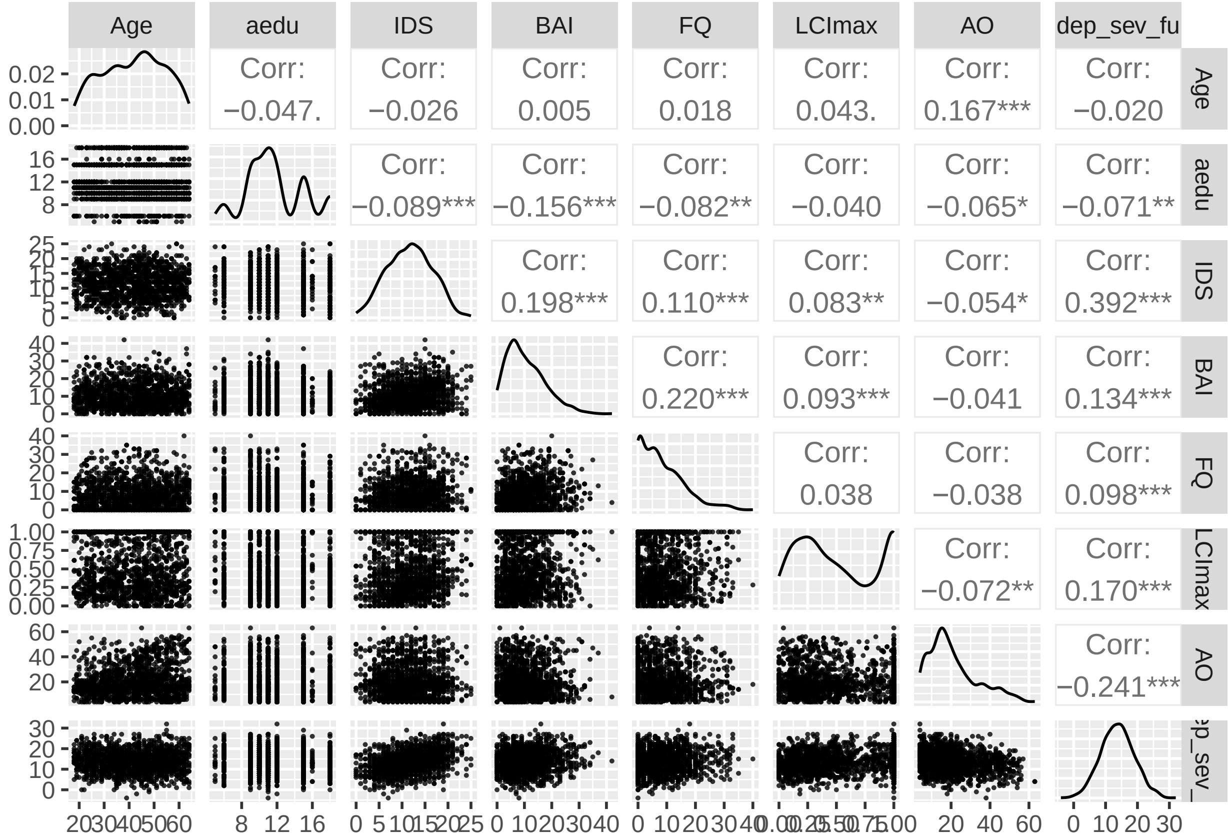
Q1 Method choice

Select three supervised learning methods from those that were discussed in weeks 9 through 11 for analyzing this dataset. Justify why you would select each of these methods for this specific prediction problem. (Use max. 200-250 words per method.)

(720 words)

Introduction

The goal of this analysis is to predict the severity of depressive symptoms at 12 months after the beginning of the study. Various characteristics of 1500 adults in the Netherlands have been measured for this purpose at the start. The target audience of this analysis are researchers in the life and behavior sciences, so models should should be interpretable/easy to use and allow for inference as well.



The given data set has 1500 observations, 20 predictors (8 numeric, 12 categorical) and 1 numeric response variable. For the training set question 2 asks us to use 1000 observations, resulting in 50 observations per predictor (ie $p < N$), so likelihood based models like GAMs can be fitted. Based on the 10:1 rule for regression models, the data set is relatively large, so high variance/low bias models like boosted ensembles should be safe to fit if tuned correctly. Regarding the irreducible error (ie noise), it is hard to tell before fitting the models as we have no additional context. But a noisy dataset would favor higher bias models like a GAM to avoid overfitting. Also, based on the uni-variate density plots above, we can observe five out of eight numeric predictors to be skewed or non-normally distributed. This could hint at potential non-linearities. Also, it might be sensible to fit models with and without interaction terms. The need for interaction terms cannot be determined beforehand.

Supervised learning method selection

I will choose two methods to focus on interpretability and inference while the third one focuses pure predictive performance to ensure a balanced approach for my target audience.

For the simplest, most interpretable method I chose the conditional inference tree model (CTree). A CTree is a good choice as researchers are presented a single tree with limited number of predictors. The tree is also a decision diagram which reads from top to bottom, so researchers can easily make a prediction without having to calculate. Additionally, researchers can identify the most important predictor at the top (root) of the tree. Predictor importance is presented in descending order as you traverse down the tree, so the least important predictors are at the bottom of the tree. I chose the CTree over a single tree fitted using cross validation (ie a pruned tree) because a CTree does not suffer from variable selection bias unlike the pruned tree method. Thus, CTrees generalise better. However, CTrees are sensitive to the data (ie are high variance/low bias) and researchers might not be happy with their predictive performance. Additionally, CTrees do not allow for inference and assume interaction terms (unless you specify a tree depth of 1).

To allow for inference, I chose a generalised additive model (GAM). Researchers might prefer a GAM over a CTree, as it has better predictive performance while allowing for inference. The performance increase comes from the additivity assumption, so no interaction terms, which leads to a more biased model that generalises better compared to a CTree. So researchers can examine the main effect of each variable while holding all other variables constant. GAMs can also be a potential high bias model when considering the data size. Additionally, researchers can infer the predictor importance via their p-values. Also, GAMs can fit non-linear relationships while still being easily interpretable. Smoothing could be relevant as the above plots hinted at non-linearities. Even though a GAM is likely the most balanced model, some researchers might prefer an uber predictive model.

For pure predictive performance I chose a boosted ensemble tree model as it is potentially the best performing model compared to the CTree and GAM, if tuned correctly. Additionally, it tends to performs best within the class of ensemble tree models. This is because boosting is a sequential tree fitting procedure, opposed to bagging and random forest which are both variations of bootstrap sampling and averaging trees. Thus, boosting tends to better capture complex, non-liner relationships. Boosting also allows for variable ranking whereas other black-box models such as SVMs do not allow for this. However, additionally to interpretability and inference issues, researchers might not like that boosting is a high variance/low bias model (even when additive, ie with tree depth of 1). So if the data is noisy then boosting, like other ensemble tree models might not perform well.

To conclude, I will use a CTree, GAM and a boosted ensemble method to ensure a balanced analysis.

Q2 Method parameters

Now apply the three methods you selected to the dataset. Beforehand, randomly split the dataset into a training (n=1000) and test (n=500) dataset. Use your student number to set the seed of the random number generator.

Motivate your choice of the main model-fitting parameters. Thus, make a well-informed choice for a fixed value of each parameter and/or use cross-validation to set their values. Your answer should reflect understanding of what each parameter does. (Use max. 200-250 words per method.)

CTree

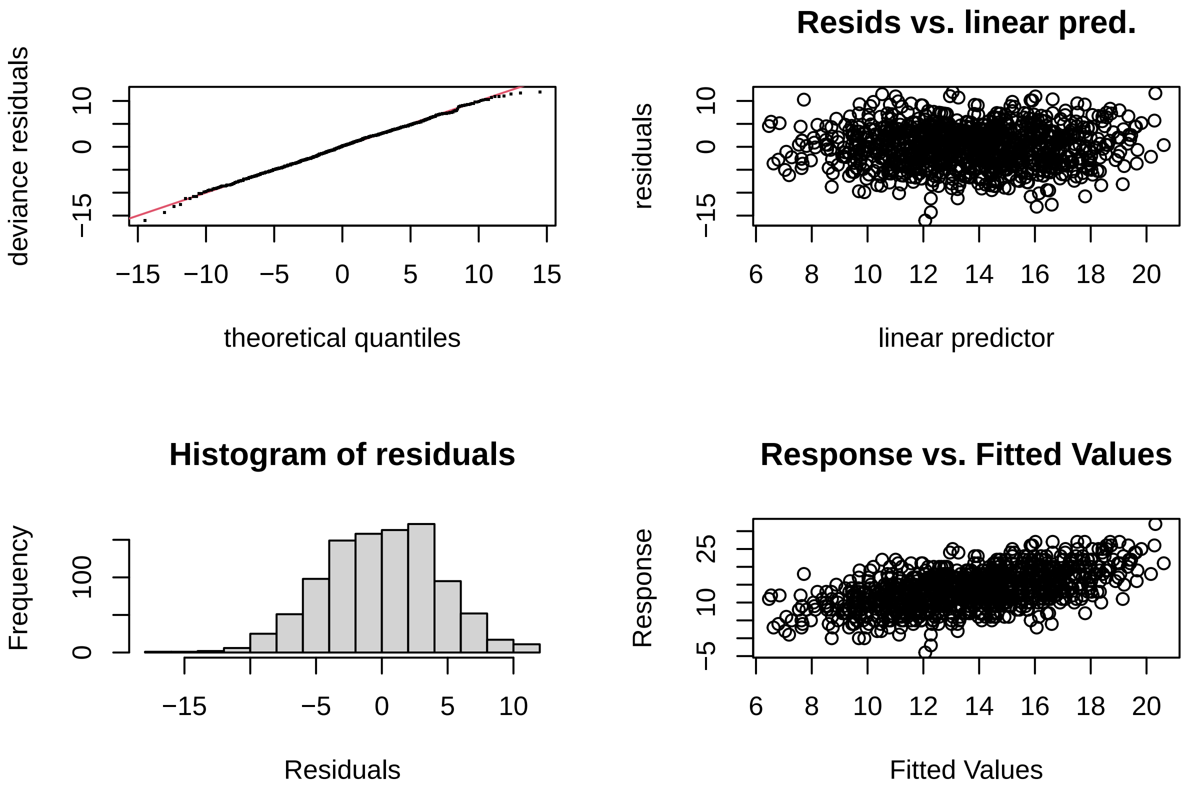
(160 words)

The main model parameter of the conditional inference tree is the chosen significance level alpha. It determines the significance level of the statistical association test used to select the splitting variable. The smaller the significance level, the harder it is for a variable to be significant, resulting in tree with fewer terminal nodes.

To fit the conditional inference tree I used 10 fold cross validation. The parameter grid only has one parameter which is the significance level. I chose three different significance levels alpha: 0.05, 0.01 and the Bonferroni corrected significance level. 0.05 and 0.01 are common significance levels. I also considered the Bonferroni correction as it adjusts for multiple testing. The Root Mean Squared Error (RMSE) was used to select the optimal model, because it gives the model performance on the same scale as the response variable. All three significance levels return similars RMSE values, but the smallest was for alpha = 0.01 which will be used for the final model.

GAM

(200 words)



The main tuning parameters of the GAM are the type of smoothing function and their smoothing coefficient, which allow the GAM to capture non-linear relationships. The smoothing coefficient, represented by the effective degrees of freedom (edf), controls the degree of smoothing. Higher edf means non-linear predictors so higher variance. Lower edf result in linear predictors, so less variance.

I used the mgcv library to identify the best parameters. It automatically selects the smoothing coefficients, but the choice of smoothing function must be specified manually. I chose the default thin plate regression spline for its ability to capture complex non-linear relationships. For boundary stability, natural splines could be an alternative.

The model fitting process began with all 20 predictors, using the REML estimation method for robust coefficient estimates. To create a parsimonious model, I performed backward selection based on AIC values, removing one variable at a time. Variables with the highest p-values were removed first, simplifying categorical variables (fixed effects) first and then numerical ones (random effects). Finally, model diagnostics (plotted above) were checked. The final model meets the inference assumptions, as confirmed by the normally distributed residuals and presence of homoscedasticity.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## dep_sev_fu ~ s(Age) + s(IDS) + s(BAI) + s(LCImax) + s(AO) + disType +
##     bSocPhob + bGAD + bAgo + ADuse + PsychTreat
##
## Parametric coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.6438     0.3325  41.030  < 2e-16 ***
## disTypecomorbid disorder     1.6525     0.3248   5.087 4.35e-07 ***
## disTypedepressive disorder   1.5951     0.3961   4.027 6.09e-05 ***
## bSocPhobPositive            -0.5500     0.3033  -1.814 0.070025 .
## bGADPositive                 1.1212     0.3264   3.435 0.000618 ***
## bAgoPositive                -0.8012     0.4539  -1.765 0.077850 .
## ADuseTRUE                   -1.3475     0.2901  -4.645 3.87e-06 ***
## PsychTreatTRUE              -0.8665     0.2815  -3.078 0.002140 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F  p-value
## s(Age)    2.7762      9  1.690 0.000597 ***
## s(IDS)    3.1647      9 15.451  < 2e-16 ***
## s(BAI)    1.6003      9  0.445 0.069230 .
## s(LCImax) 0.9754      9  1.772 4.17e-05 ***
## s(AO)     1.4727      9  6.460  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.273   Deviance explained = 28.5%
## -REML = 2912.5  Scale est. = 19.361    n = 1000
```

The final GAM model includes 6 categorical and 5 numeric variables, with smoothing coefficients indicated by the edf column.

Boosting

(250 words)

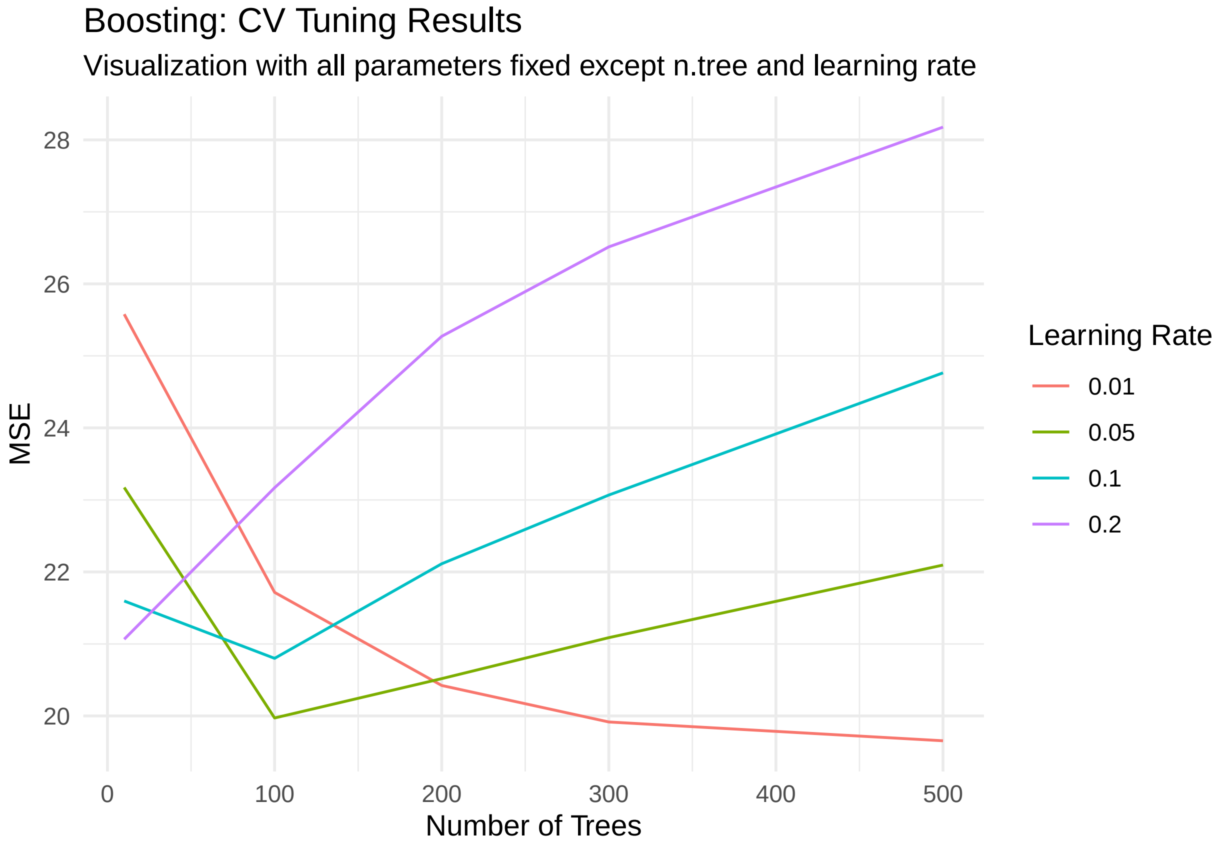
Boosting has many parameters that can be tuned. However, I am following the advice given during lectures and will limit myself to the following three: Total number of trees (B), learning rate λ and maximum number of terminal nodes per tree.

As previously explained, boosting is an iterative process and the total number of trees is the upper limit. Unlike bagging and random forests, boosting can overfit if B is too large and underfit if B is too small. The learning rate λ controls how fast the boosting model learns. Small λ can require larger B values for the model to converge, but usually result in better performance. The maximum number of terminal nodes controls if the boosted ensemble is fitting an additive model or not. A tree with only two terminal nodes has only one splitting rule, so it can be interpreted as a main effect. Any tree with more than one splitting rule captures interaction effects via its sequence of nodes until the terminal. A boosted ensemble with interaction terms is higher variance than an additive boosted ensemble.

To identify the best model parameters I used 10 fold cross validation (CV) on the training set.

Here are my final boosting parameters. The parameter `n.minobsinnode` had to be included for the CV function to work but was kept fixed at its default value and can be ignored.

```
##  n.trees interaction.depth shrinkage n.minobsinnode
##      500                 5      0.01             10
```



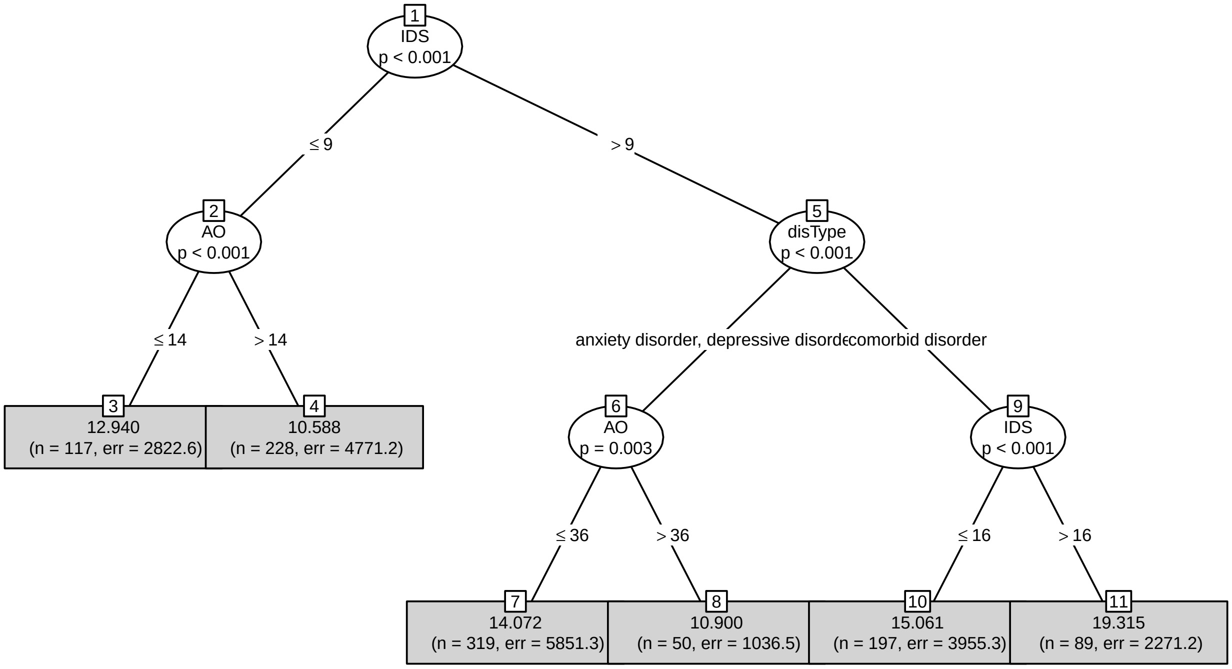
This plot visualizes how smaller learning rates tend to perform better but require more trees, indicating their trade off relationship.

Q3 Results & predictors

Provide an interpretation of each of the resulting models: Describe which variables are most important in determining the value of the outcome variable, and which measure(s) you used to determine their relative importance. Describe the effect of the most important variables (e.g., describe the shape and direction of the effect on the outcome and/or provide and discuss plots of the variables' effects) for each method.(Use max. 150-200 words per method.)

CTree

(200 words)



Based on the above conditional inference tree, we can see that the following three variables are the most important in determining the response variable:

1. Inventory of Depressive Symptomatology (IDS) score of the patient
2. The patients type of disorder (disType)
3. Their age at onset of the disorder (AO)

The patients IDS score is the most important variable as it sits at the root node. Then the disorder type which is the second node and finally the onset age.

The importance of these variables is inherent to the conditional inference tree's structure. It is computed via statistical association tests conducted at each split, meaning that variables closer to the root have a stronger effect on the response variable.

The model can be interpreted as follows. Patients with an IDS score of 9 points or lower, with an onset age of older than 14, will have the lowest depressive symptoms at 12 months post study begin. Their response variable is expected to be 10.6 points. The highest expected response is 19.3 points which is expected by patients with an IDS score of higher than 16 who have comorbid (both anxiety and depressive) disorder. Notably, the model splits the disorder type into comorbid versus single-disorder (either anxiety or depressive) categories.

GAM

(200 words)

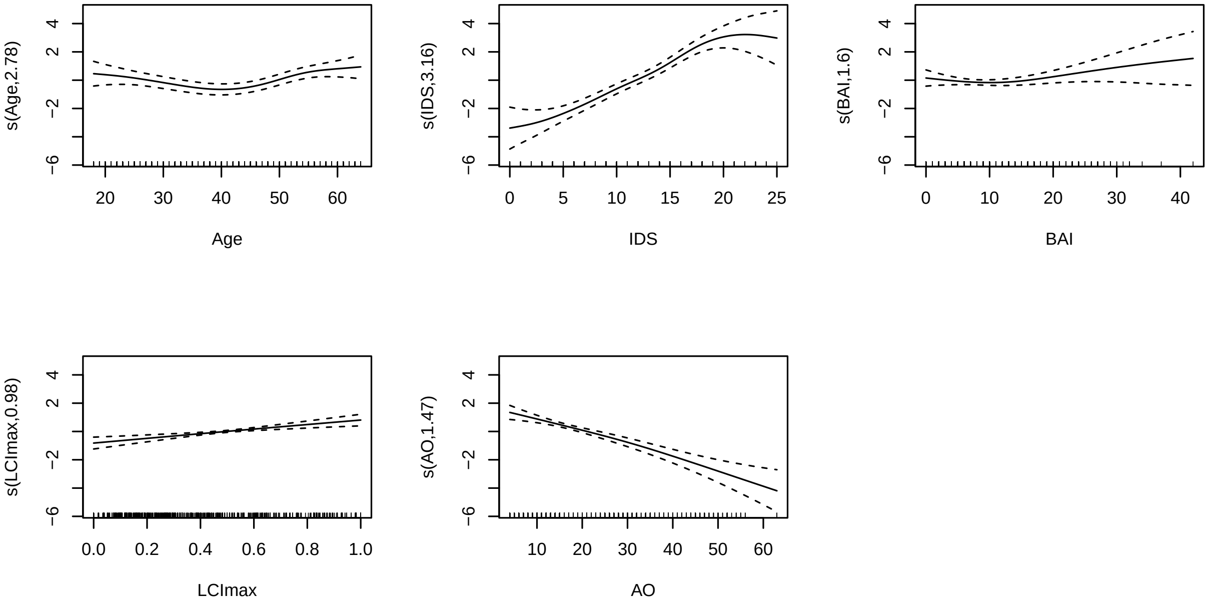
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## dep_sev_fu ~ s(Age) + s(IDS) + s(BAI) + s(LCImax) + s(AO) + disType +
##     bSocPhob + bGAD + bAgo + ADuse + PsychTreat
##
## Parametric coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.6438     0.3325  41.030  < 2e-16 ***
## disTypecomorbid disorder     1.6525     0.3248   5.087 4.35e-07 ***
## disTypedepressive disorder   1.5951     0.3961   4.027 6.09e-05 ***
## bSocPhobPositive            -0.5500     0.3033  -1.814 0.070025 .
## bGADPositive                 1.1212     0.3264   3.435 0.000618 ***
## bAgoPositive                -0.8012     0.4539  -1.765 0.077850 .
## ADuseTRUE                   -1.3475     0.2901  -4.645 3.87e-06 ***
## PsychTreatTRUE              -0.8665     0.2815  -3.078 0.002140 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F  p-value
## s(Age)    2.7762      9  1.690 0.000597 ***
## s(IDS)    3.1647      9 15.451  < 2e-16 ***
## s(BAI)    1.6003      9  0.445 0.069230 .
## s(LCImax) 0.9754      9  1.772 4.17e-05 ***
## s(AO)     1.4727      9  6.460  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.273   Deviance explained = 28.5%
## -REML = 2912.5  Scale est. = 19.361    n = 1000
```

For the GAM, we identify the most important predictors by examining p-values. T-tests for categorical variables and F-tests for numerical variables. A smaller p-value indicates higher importance.

For categorical variables, the type of disorder is most important, followed by anti-depressant use. With anxiety as the reference, having comorbid disorder increases the response by 1.65 points, while depressive disorder increases it by 1.6 points. Anti-depressant use improves the outcome by decreasing the response by 1.35 points. Categorical variables shift the regression intercept but do not affect the slope.

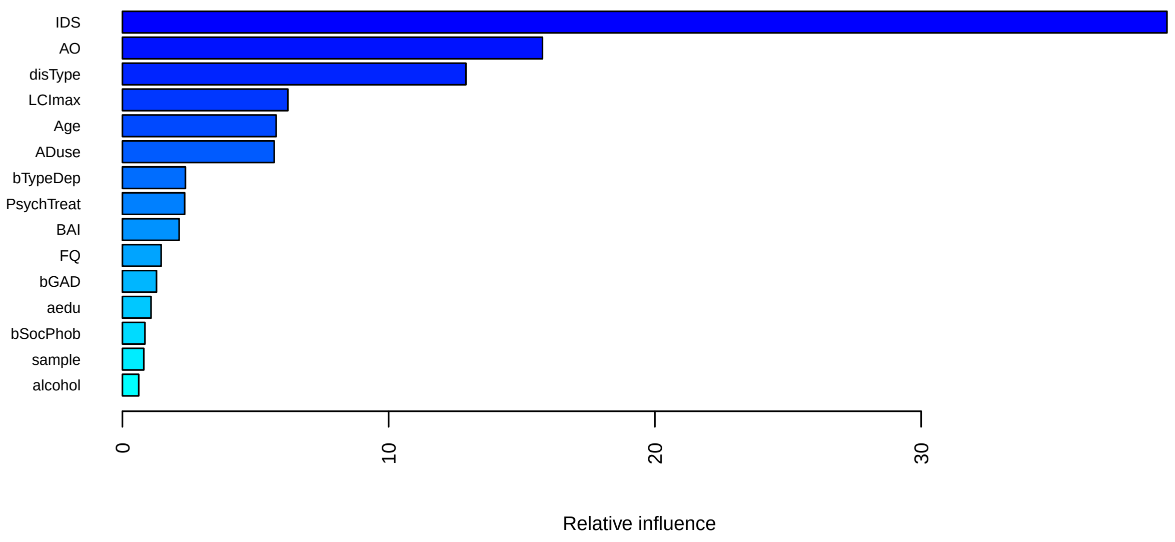
Among numerical variables, the patient's IDS score is most significant, followed by onset age and then "time with disorder symptoms" (LCImax). A 5-point increase in IDS score leads approximately to a 2-point increase in the response, indicating a direct linear slope. Conversely, a 10-year increase in onset age reduces the response by approximately 1 point, suggesting an inverse linear slope. Boundary behavior is ignored as it is less reliable. All other variables are held fixed when examining slopes.

The plots illustrate these conditional effects with 95% confidence intervals, showing the variable direction within their interquartile ranges. Conditional effects allows for negative values even though the response variable is non-negative.



Boosting

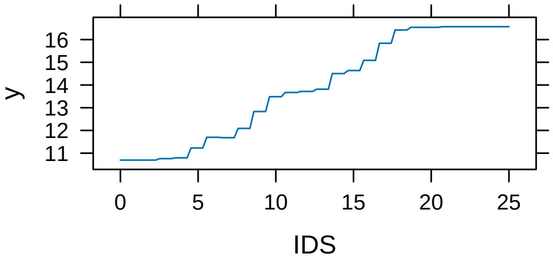
(200 words)

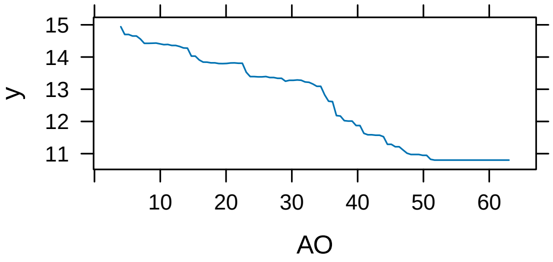


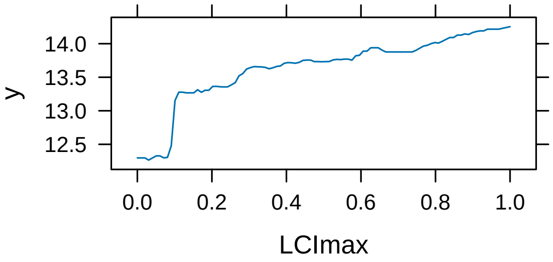
Boosting ensembles cannot be directly interpreted like GAMs or single trees. However, their variables can be ranked by importance. The above plot shows each variable's relative influence in reducing the model's loss function, measured using a permutation test on the full training set. This test estimates importance by shuffling features and measuring the increase in prediction error.

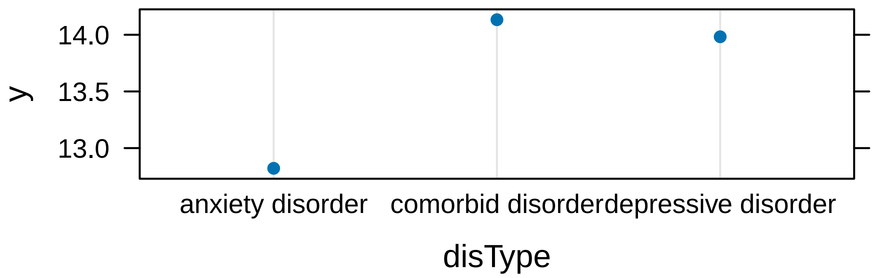
Similarly to the previous two methods, the most important variable is the IDS score, followed by the onset age, then disorder type. The "time with disorder symptoms" (LCImax) variable can be considered the best of the (unimportant) rest.

Partial dependence plots (PDPs) visualize the effects of the most important variables, showing the effect of a single predictor averaged over all other predictors The y-axis is on the original scale as the response variable. IDS shows a linear trend with a 10 point IDS increase leading to roughly a 2 point increase in the response. Onset age shows an inverse linear trend with a 10 year increase in onset age resulting in approximately a 1 point decrease in the response. The disorder type is categorical and comorbid disorder has the highest response value, so the worst outcome. LCImax shows a linear trend with a flatter slope, indicating a less pronounced effect.









Q4 Predictive accuracy

Assess and compare the predictive accuracy of each of the models using the test set. Which model predicts best? Bonus: Using a suitable approach, compute confidence intervals for the (pairwise differences in) predictive performance (not taught during the lectures). (Use max. 100 words, max. 200 words including bonus.)

CTree

```
## [1] "Ctree: Test set MSE is  21.5658  and RMSE is  4.6439"
```

GAM

```
## [1] "GAM: Test set MSE is  19.2487  and RMSE is 4.3873"
```

Boosting

```
## [1] "Boosting: Test set MSE is  19.1101  and RMSE is  4.3715"
```

Evaluation

(50 words)

As shown above, the boosting ensemble has the lowest test set MSE, indicating the highest predictive accuracy. It is thus the best predicting model, closely followed by GAM. CTree exhibits the poorest performance. Boosting and GAM have similar accuracies (rounded RMSE of 4.4), while CTree lags significantly.

Bonus Evaluation

(150 words)

The pairwise difference in predictive performance with 95% confidence interval is given below:

```
##            Models Confidence_Interval Significant.Pairwise.Difference
## 1   GAM vs CTree  [-5.5119, -0.6018]                     Significant
## 2   GAM vs Boost   [-1.7975, 0.6105]                 Not significant
## 3 CTree vs Boost    [0.3566, 4.5097]                     Significant
```

To compute this, I locked the test away only considered the training set, call it MC set. I repeatedly fitted the three models on a randomly generated training subset of the MC set. The three models used the best tuned parameters identified in Q2. By fitting the best tuned models on the same subset I ensure that the models are comparable. I then evaluated their MSE based on the training subset of the MC set. This process was repeated 1000 times, and empirical percentiles were used to compute confidence intervals.

```
##   Models Mean_MSE Confidence_Interval
## 1    GAM 19.21839  [15.9788, 22.7731]
## 2  CTree 22.38266  [18.3693, 26.3321]
## 3  Boost 19.81276  [16.6163, 23.1716]
```

Based on this simulation, the on average best performing model is now the GAM instead of boosting. However, since their difference in accuracy is not significant either model can be seen as the "best predicting" model depending on the problem context. The CTree is consistently inferior. All test set MSE's fall inside the empirical confidence intervals.

Q5 Conclusion on predictors

Based on 3 and 4: Provide a short overall conclusion regarding which predictors are related to the outcome. (Use max. 100 words.)

(100 words)

Based on question 3, all three models agree that a patient's IDS score is the most important predictor. Both the CTree and the boosting ensemble agree that onset age is the second most important predictor, followed by the disorder type. The GAM does not clearly rank between onset age (numerical) and disorder type (categorical) but includes both among the top predictors.

Based on question 4, we know that boosting and the GAM perform significantly better than the CTree, partially explained by including additional variables. Thus, the "time with disorder symptoms" (LCImax) variable can *potentially* be considered the fourth most important predictor. All other variables seem unimportant.

Q6 Individual prediction

A psychologist has seen David Edgar Pression for an intake today. The psychologist wonders whether they should refer David to an intensive depression treatment program.

The psychologist asks you to provide them with an estimate of the severity of David's depressive symptoms in 12 months. Patients with predicted depressive symptom severity equal to or greater than 17 are referred to the intensive treatment program. What is your estimate? Should David be referred to the intensive treatment program? Bonus: Using a suitable approach, quantify the uncertainty of your estimate (not specifically taught during the lectures). (Use max. 100 words, max. 200 words including bonus.)

CTree

```
## [1] "Ctree: David's predicted depressive symptoms severity in 12 months is  14.0721"
```

GAM

```
## GAM: David's predicted depressive symptoms severity in 12 months is  18.6141
```

```
## With 95% Confidence Interval: [ 17.4355 ,  19.7927 ]
```

Boosting

```
## [1] "Boosting: David's predicted depressive symptoms severity in 12 months is  16.8054"
```

Evaluation

(200 words)

As shown above the three models results in three different estimates, even when rounded to the nearest integer. From question 4 we know that the CTree on average predicts significantly worse than both other models. Thus, I will ignore the CTree prediction. Surprisingly, the ensemble and the GAM return contradicting predictions, differing by more than 2 points on the response variable scale.

Examining the top three numerical predictors reveals that the top two (IDS score and onset age) are on the boundary of their variable space. This might explain the discrepancy between GAM and boosting as predictions are least stable at the boundaries.

```
## David's Inventory of Depressive Symptomatology (IDS) score
##  is in the 92.2 th percentile of all data points in the training set
```

```
## David's Age at onset (AO)
##  is in the 3.5 th percentile of all data points in the training set
```

```
## David's Percentage of time in which symptoms of anxiety
##  and/or depressive disorders were present during the past four years
##  (LCImax) score is in the 68.8 th percentile of all data points in the training set
```

As opposed to tree based models, the GAM is likelihood based. We also know from the model diagnostics in Q2 that the GAM meets the inference assumptions. Thus, the GAM can quantify the uncertainty around its predictions via standard error calculations. Q4 indicates that the GAM's predictive performance is on par with the boosting ensemble model, while being more interpretable and allowing for inference. Thus, I will select the GAM's estimate.

My final estimate for David's depressive symptoms is 18.6141. Since the 95% confidence interval (17.4355, 19.7927) is above the threshold value of 17, David should be referred to the intensive treatment program.