Statistical Learning - Individual Assignment 02

Joshua Damm

2024-05-21

Question 1: Model selection and reasoning

1.1 GAM

One reasonable supervised learning method to predict the severity of depressive symptoms after 12 months (from now on: depressive symptoms or depression score) could be generalized additive models (GAM's). We have a continuous response, and a range of many categorical (many of them even binary) and continuous predictors. GAM's allow to model non-linear relationships for seperate predictor variables with the response, while still remaining the additivity assumption of a typical linear model, so that one can interpret each predictor's effect on the reponse seperately, while holding all other predictors constant. It is therefore possible to include different effects (both linear and non-linear) for higher predictive accuracy. Since our predictors have different scales and might have both linear and non-linear relationships with the response, it might be reasonable to fit a GAM to the data.

1.2 GBM

Another way to predict depression score could be using regression tree ensembles. We have a large epidemiological dataset with potentially complex non-linear relationships and interactions among the (bio-)psychological predictor variables, and regression trees (continuous response) are paticularly good in dealing with this. One advantage of using tree methods is that one can infer the role of different predictors for the prediction by looking at the shape and splits of the trees that were used to terminate at a prediction for the response.

Gradient Boosting Machines (GBM) are particularly suited for predicting depression score since multiple trees will be fitted sequentially on our dataset, where each new tree helps to correct errors made by previous trained trees. This additive model is expected to perform particularly well on this heterogeneous dataset and can capture its potentially complex interaction patterns. By fitting small trees, and averaging predictions over many trees, both overall bias and variance will be reduced. This process is regulated by different tuning parameters like the learning rate (or shrinkage parameter), number of trees and number of splits that allows for different shaped trees to attack the residuals. One other advantage is that by the number of splits the interaction depth from the variables predicting the response can be controlled. Clinicians could for instance infer useful information by looking at the most influencial variables that reduced the RSS among the different splits in the training process the most, to potentially derive theories or interventions.

1.3 Single regression tree

Another suitable method to predict later depression score would be to use a single (pruned) regression tree. Firstly, regression trees provide a straightforward and intuitive way to visualize and interpret how our biopsychological predictor variables influence the reponse The tree structure, with its branches and leaves, allows for easy understanding and communication of the model's decision process, which is particularly beneficial

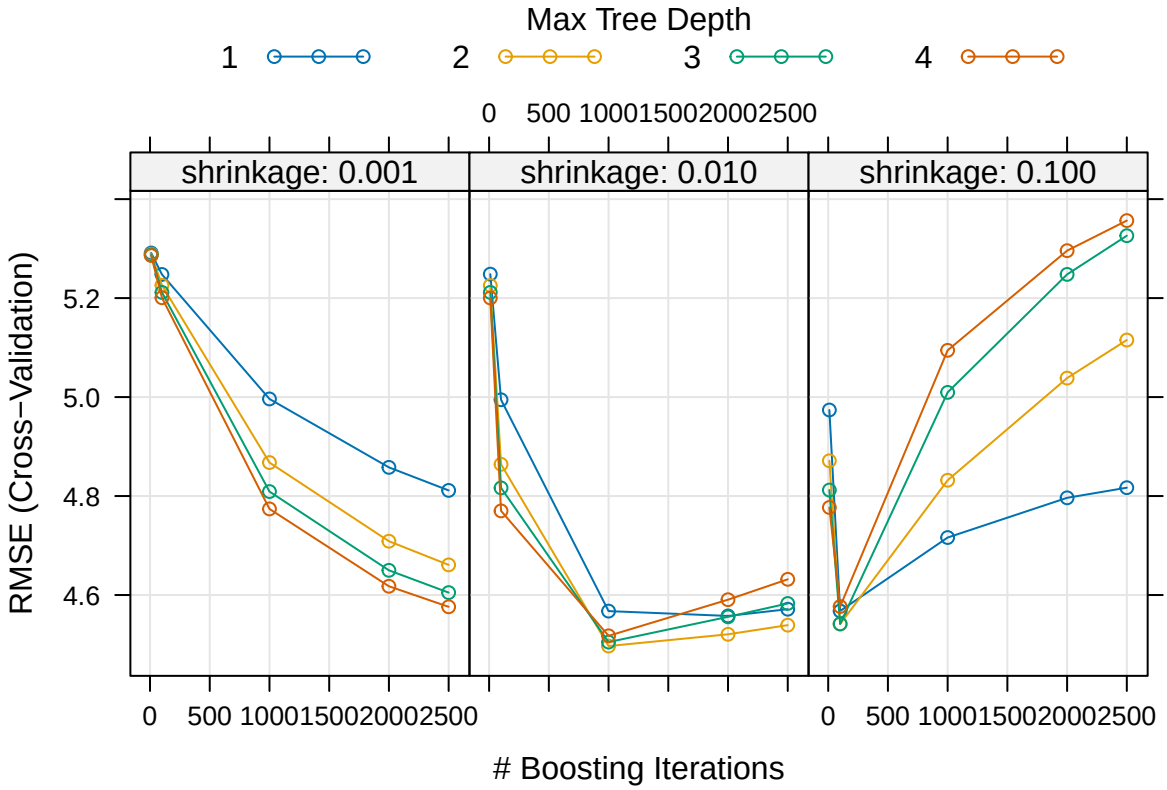
in a medical and psychological context where clear explanations are crucial. Secondly, regression trees can naturally handle non-linear relationships between predictors and the outcome without requiring transformation of variables. Additionally, they can capture interactions between variables, which is often essential in complex datasets. For instance, the interaction between different types of disorders and demographic factors might significantly affect depression severity, and regression trees can model these interactions effectively. Regression trees perform implicit variable selection, identifying the most relevant predictors for the outcome. This is advantageous for reducing dimensionality and focusing on the most impactful variables, which can enhance model performance and interpretability. Pruning further refines the tree by removing less significant branches, preventing overfitting and improving generalization to new data.

Question 2: model fitting and parameter selection

2.1. GAM

For the GAM I used all categorical variables as parametric predictors and all continuous predictors as smoothing splines. One should use cross-validation to determine the most effective choice of degrees of freedom (df) for the smoothing splines. However, the "mgcv" package already uses generalized cross-validation (GCV) as the default method to select the amount of smoothing, which is a form of automatic smoothing parameter selection that's closely related to cross-validation. The package also provides a method to choose the model parameters according to restricted maximum likelihood estimation (REML). In this case the df are not specified, but the REML estimates the optimal value for the smoothing parameters on the cross-validated training data. So for the smoothing splines the df were chosen using GCV and REML. All categorial variables were included as factors in the model.

2.2 GBM

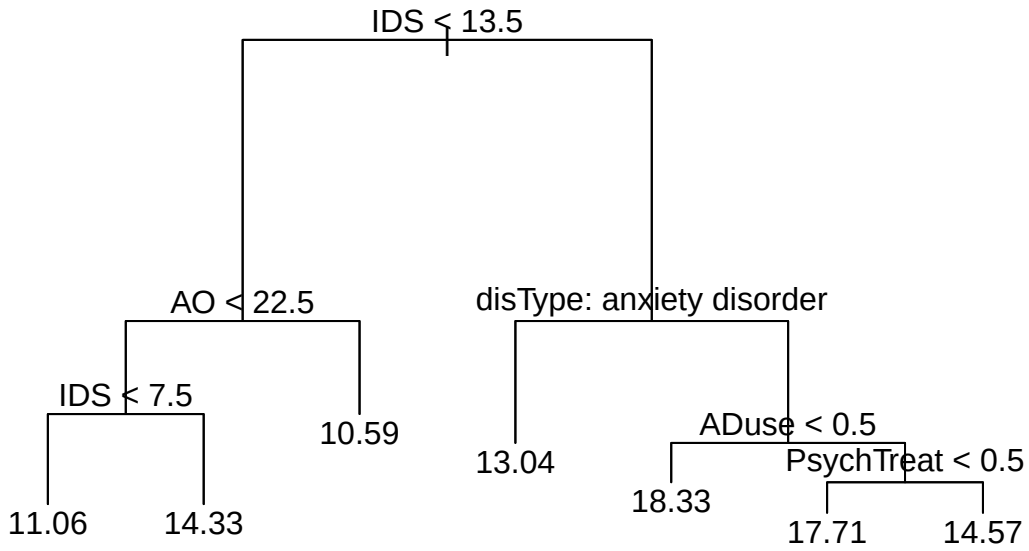


I used 10-fold cross validation to choose the hyperparameters of the gradient boosted regression model to predict depression score. I defined a grid including the different hyperparameters, i.e. the number of trees, the learning rate, and the tree depth.

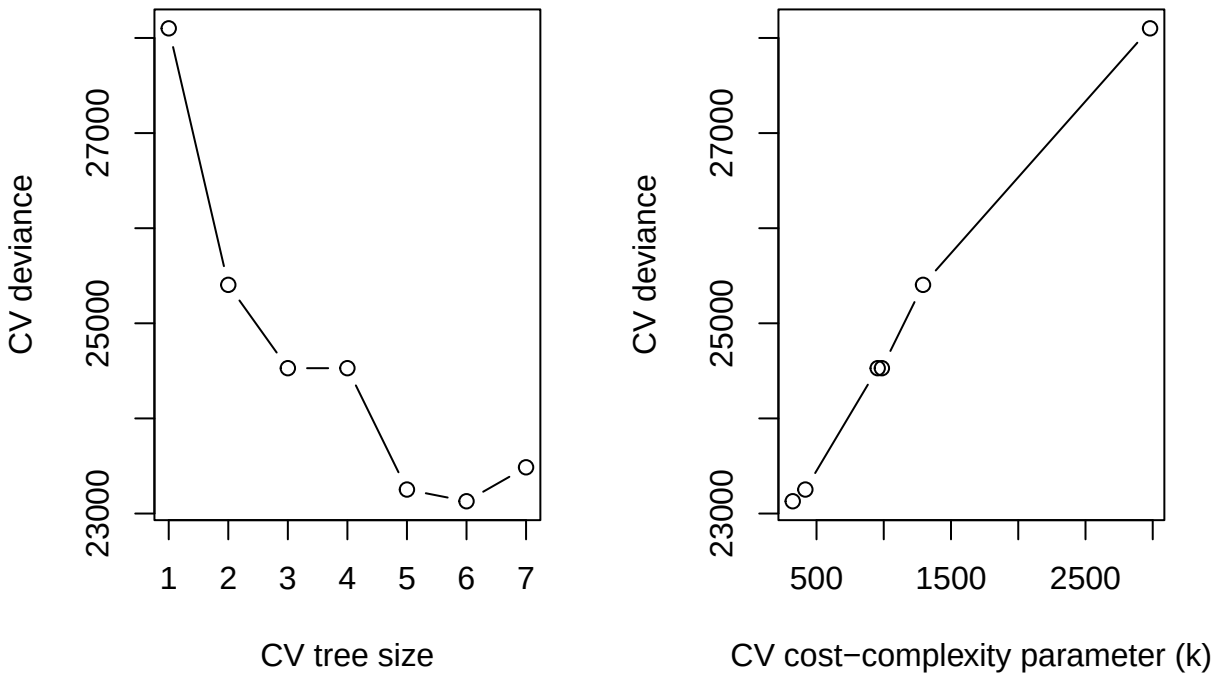
The number of trees parameter specifies the total number of trees to build in the sequence. Each tree is built on the residuals (errors) left by the previous trees. More trees can lead to better performance but increase the risk of overfitting. The learning rate (shrinkage parameter) used to shrink the corrections made by each tree, thereby slowing down the learning process. It is used to prevent overfitting by making the boosting process more conservative. The tree depth pecifies the maximum (interaction) depth of each tree. Deeper trees can model more complex patterns in the data. More depth allows the model to learn more detailed data specifics (for instance interaction of multiple predictors), enhancing performance on training data but can lead to overfitting. The minimum observations in Leaf Nodes controls the minimum number of samples required to be at a leaf node of a tree, which also contributes to the bias-variance trade-off.

After inspected the Root Mean Squared Error (RMSE) after cross validation (see plot), a number of 1000 trees, a learning rate of 0.01, an interaction depth of 2 showed the best model performance. This minimum observations in Leaf Nodes was fixed as a value of 10 as this is conventionally a good strategic choice aimed at enhancing the model's generalizabilty and preventing overfitting.

2.3 Single regression tree



First I fitted a single regression tree on the training data with later depression score as the continuous response. The regression tree split the data onto different segments based on decision points (nodes). The first optimal choice of splits / nodes was based on minimizing the total residual sum of squares (RSS, also called deviance in the context of regression trees), leading to a total of 6 splits / nodes and 7 terminal nodes (i.e. end points of the tree for final predicted response values). After this I used cross validation to see if we can prune the tree (i.e. shorten the branches and nodes of the tree) to make it less complex and to introduce a little bias for the potential reduction of variance and overfitting. I investigated the cross-validated deviance as a function of cross-validated tree size and the cross-validated cost-complexity parameter (k). We want to minimize the deviance for reducing the bias, while also accounting for the variance and the risk of overfitting, thereby introducing the cost-complexity parameter k. Initially, the full-grown tree typically overfits the training data. The cost-complexity criterion introduces a penalty for the number of terminal nodes in the tree. The optimal subtree that the fewest necessary complexity while having the lowest cross-validated error (deviance), turned out to be k = 323.2192, with a tree size of 6 and a deviance of 23486.26 (see plot).



Question 3: Interpretation of the results

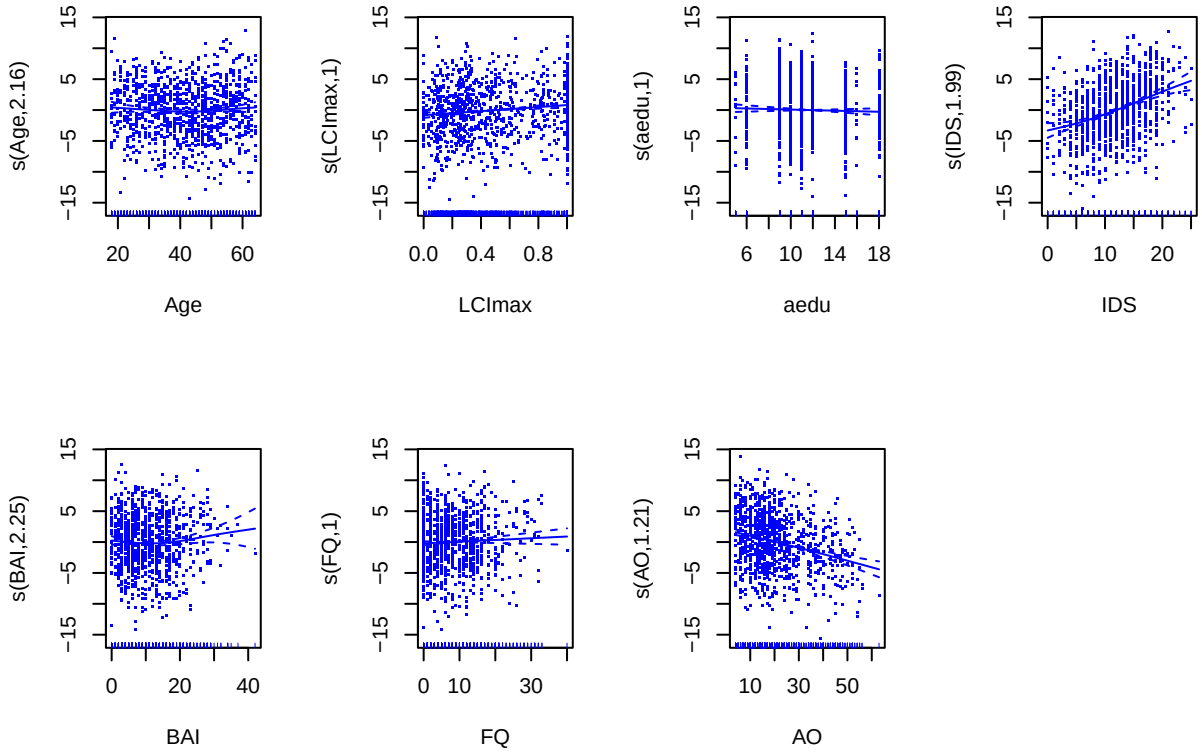
3.1 GAM

The model showed that the type of disorder matters ($F = 12.156$, $p < .0001$). Having a depressive disorder is predictive, and having a comorbid disorder is even higher predictive of depression score than having only an anxiety disorder. Comorbidity matters a lot, since having a social phobia seems is negatively associated with depression score ($F = 4.037$, $p < .05$), whereas having a general anxiety disorder is positively associated ($F = 10.091$, $p < .01$) with the response. While having a panic disorder is not related to the response, having an agoraphobia is negatively associated with depression score ($F = 5.986$, $p < .05$). Taking antidepressive medication ($F = 25.911$, $p < .00001$) and receiving psychological treatment is both negatively associated with depression score ($F = 15.373$, $p < .00001$).

Furthermore, when looking at the smoothing spline plots visually, the percentage of time of present symptoms in the past 4 years seems to have a slightly positive association with depression score ($F = 14.954$, $p < 0.001$), while the age of disorder onset has a rather strongly negative associaton with the response ($F = 49.782$, $p < 0.00001$). While the former two effects are rather linear (edf value close to 1), the test score on the inventory of depressive symptomatology has a strong rather non-linear relationship with the reponse (edf value close to 2).

Overall, the F-test statistic suggest that the most influential variables are the age at onset of the disorder, the test score on the inventory of depressive symptomatology, whether subject uses anti-depressant medication, whether a subject gets psychological treatment, the percentage of time of present symptoms in the past 4 years, the type of disorder, and the presence of a (comorbid) anxiety disorder, in descending order. This is

an arbitrary selection, since there would be 2 more significant predictor variables with substantially lower F-values.

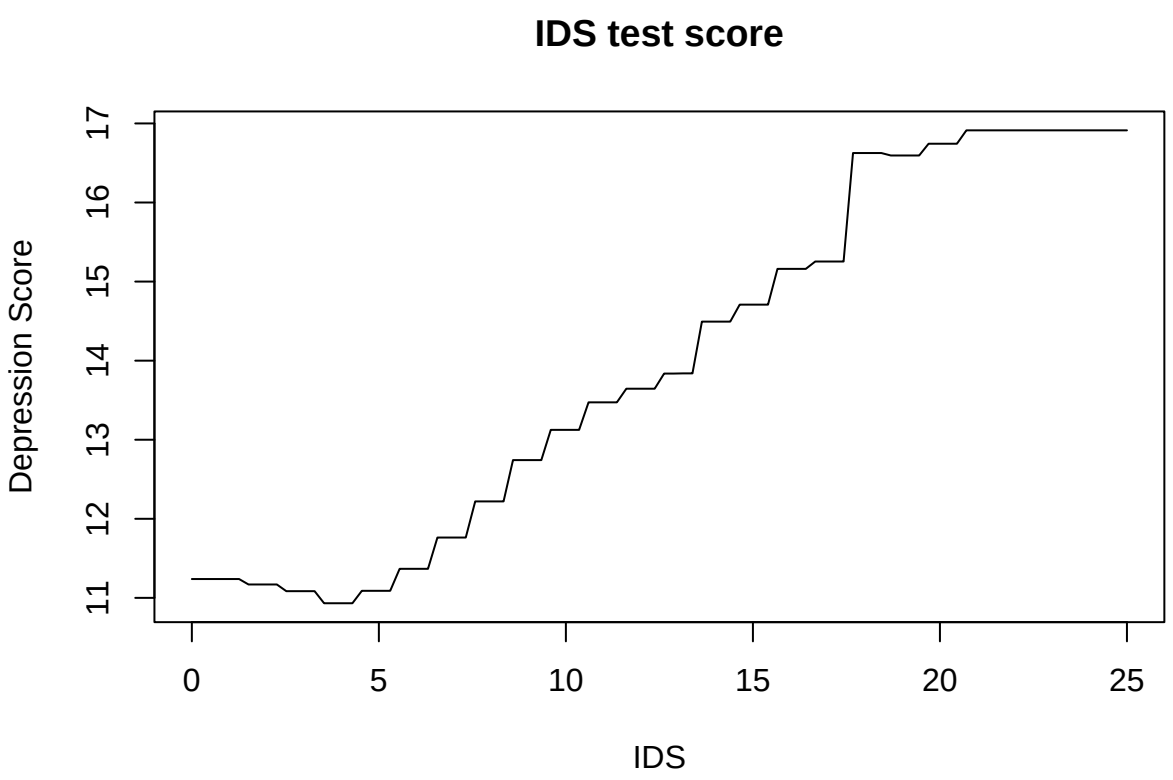


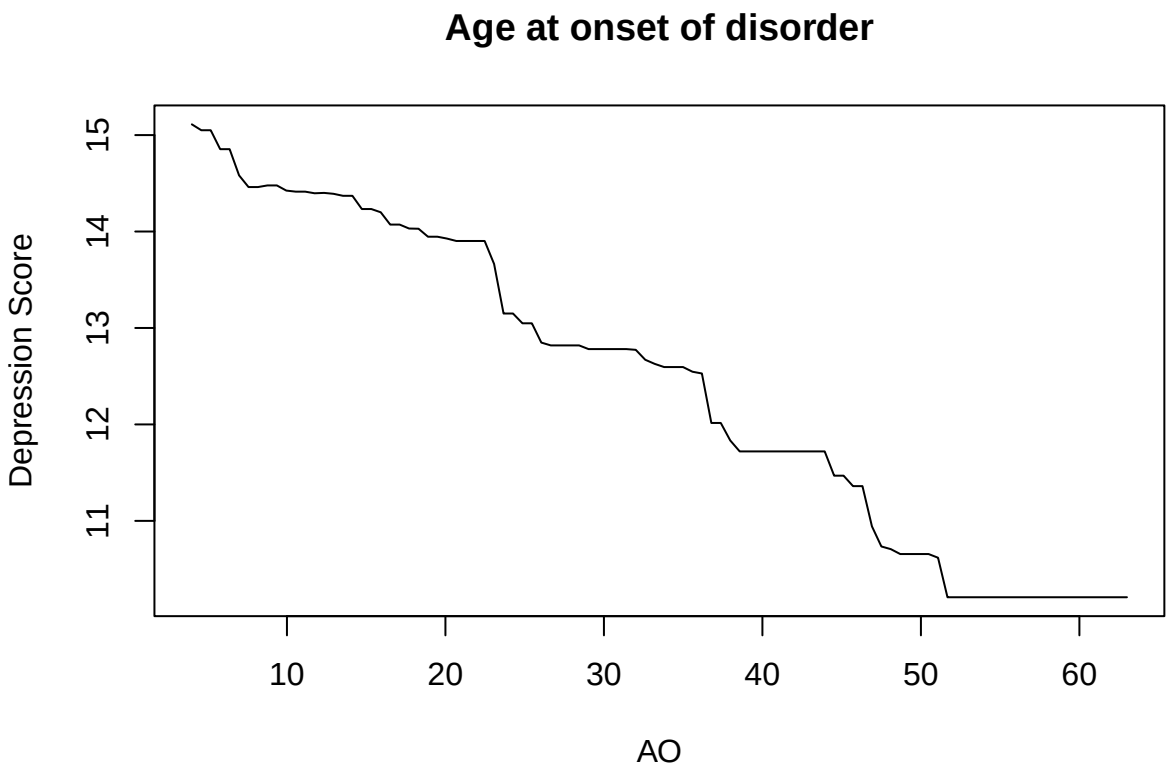
3.2 GBM

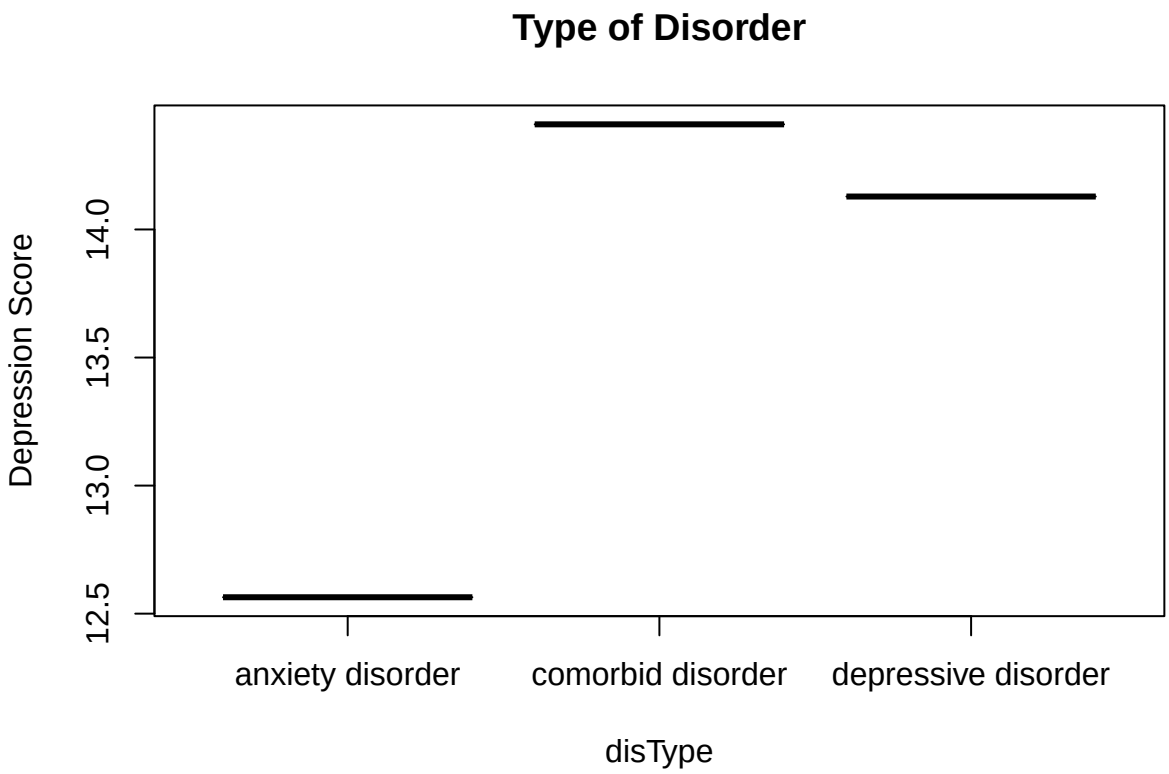
To see which variables were most important in predicting depression score, I inspected the relative importance of variables by looking at the normalized total gains across the tree-building/learning process. Each time a feature is used in a split, the improvement to the model (measured by the reduction in the loss function, e.g. MSE) is calculated. The total gain from each feature across all splits it participates in is aggregated and can be used as a measure for variable importance. The scores are then normalized for all features so that they app to 1 or 100%.

Looking at variable importance, we see that the test score of the Inventory of Depressive Symptomatology (IDS) is the most influencial variable (31.3%), followed by Age at onset of the disorder (AO) (16.04%) and Type of disorder (disType) (12.67%).

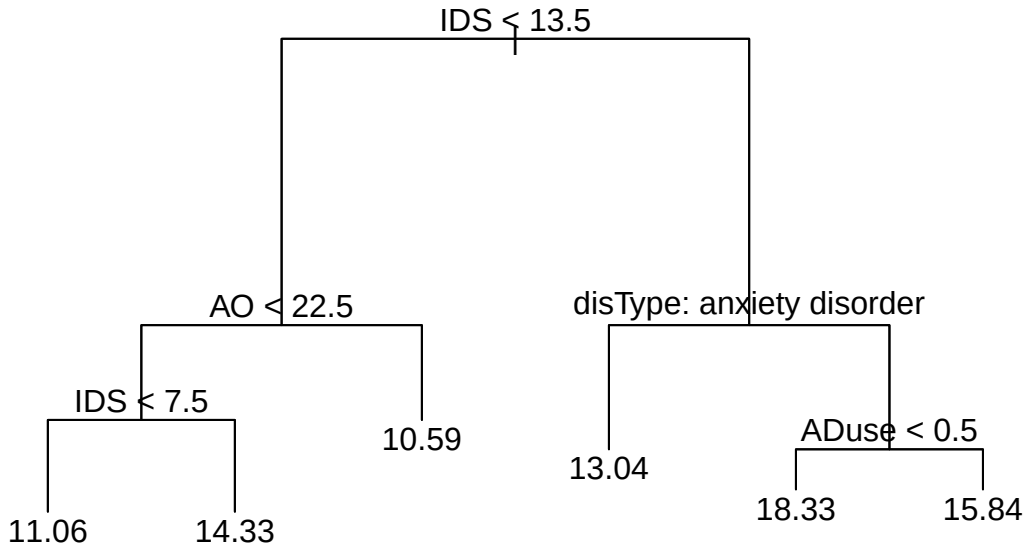
When looking at the shape and direction of the effects of the most important variables, one can use partial dependence functions, which compute the conditional expectations for the respective predictor variables of interest by marginalizing over the observed distributions of all other predictors (see plots). The Inventory of Depressive Symptomatology (IDS) seems to have a positive association with depression score. Especially higher values from about 16 seem to be increasingly predictive of later depression score. The later one is affected by the disorder, the lower the depression score at the follow-up, as indicated by the negative association between age at onset of the disorder and the response. Comordbid and depressive disorder types are more predictive of later depression score than only having an anxiety disorder.


IDS test score
Depression Score
11
12
13
14
15
16
17
0
5
10
15
20
25
IDS


Age at onset of disorder
Depression Score
15
14
13
12
11
10
20
30
40
50
60
AO


Type of Disorder
Depression Score
12.5
13.0
13.5
14.0
anxiety disorder
comorbid disorder
depressive disorder
disType

3.3 Single (pruned) regression tree



To see which variables were most influential in predicting later depression score when using a single regression tree, one can visually inspect which variables were used at which stage for splitting. Candidate variables that were used for earlier splits showed less deviance in predicting the response than other candidates. The variables that were used in tree construction were the Testscore on the Inventory of Depressive Symptomatology (IDS), Age at onset of disorder (AO), Type of Disorder (DisType), and wether a subject uses anti-depressant medication (ADuse).

As we can see the first split was driven by the the Testscore on the Inventory of Depressive Symptomatology (IDS). Higher IDS scores seem to predict higher later depression score. For patients with lower IDS scores, the age at the onset of the disorder seems to play a crucial role. Patients who were younger than 22.5 years at the onset of the disoder and have a higher IDS score than 7.5 show substantially higher later depression score than people the same young patients with an IDS score lower than 7.5 or patients older than 22.5 years at the onset of the disorder.

Furthermore, patients with an IDS score higher than 13.5, and who were diagnosed with an anxiety disorder, show the highest depression scores. From the patients, only patients that do not use antidepressant medication show the highest predicted later depression scores.

Question 4: Predictive accuracy, confidence intervals

I used the Mean Squared Error (MSE), and the R^2 to compare each model's prediction performance against each other on the test data. The R2 is a measure of how much variation in the response is explained by the model's predictions. The Generalized Additive Model had an MSE of 18.25573 and could capture 26.36% of the variation in the response in the test data. The Gradient Boosted Regression Machine achieved an

MSE of 18.41213 and could capture 25.73% variation in the response. The single regression tree had an MSE of 22.34438 and captured only 9.87% variation in the response in the test data. Therefore, one can conclude that all models show poor predictions, but the single pruned regression tree performed worst. The best prediction was achieved by the generative additive model.

BONUS: Confidence intervals for predictions

To compute the confidence intervals for the pairwise differences in predictive performance I generated a bootstrapped sample of the test data for the response variable and the predictions for the reponse by the respective models, calculated the MSE for each bootstrapped sample (sampling with replacement), computed the pairwise MSE differences between the models and constructed confidence intervals (CI) from the distribution of the bootstrapped MSE differences (using the 2.5th and 97.5th percentiles of a 95% CI). The Confidence Interval for the MSE difference between GAM and GBM is [-0.7603545, 0.4164156]. The Confidence Interval for the MSE difference between GAM and the single tree is [-5.800563, -2.298901]. The Confidence Interval for the MSE difference between GBM and the single tree is [-5.448023, -2.477194].

Question 5: Conclusion of most influential variabels

To see which predictor variables are related to later depression score, we can investigate which predictor variables were influential across all methods used. Overall, taking into account all models, the Inventory of Depressive Symptomatology (IDS), the age at onset of the disoder (AO), the type of disorder (DisType), and whether subject uses anti-depressant medication (ADuse). This is again an arbitrary consideration since it depends on the importance metrics and architectures of the respective models. There were also other inluential variables detected by single models like whether a subject gets psychological treatment (PsychTreat), which were not found in other models.

Question 6: predict patient depression score

To provide an estimate of the severity of David's depressive symptoms after 12 months I used the previously fitted models and provided David's data as the test data to predict his later depression score. The GAM predicted a score of 17.7, the GBM predicted a score of 17.39, and the single regression tree predicted a score of 13.04. So according to 2 out of 3 models, we should send David to the intense Depression treatment program. According to the single regression tree we should not send David to the program. However, the single regression tree also showed very poor performance based on the MSE and the R^2, which could be a reason to weigh its prediction less. We can also look at the uncertainty of our estimate for the final decision.

BONUS: Quanitfy uncertainty of the estimate

I constructed confidence intervals for the predicted values based on the respective standard errors from the cross-validated model fitting processes. For the GAM the 95% prediction interval lies between (16.29379, 19.11403). For the GBM the 95% prediction interval lies between (9.284026, 25.505682). For the single regression tree the 95% prediction interval lies between (3.758488, 22.312308). Based on the width of the model's prediction intervals we can see that all prediction intervals include values both above and below the threshold of the referred treatment program. While for the GAM the threshold lies on the lower end of the interval, favoring a decision for the program, the intervals for GBM and the single tree span too far on both ends, signaling a high uncertainty of the estimate. So for the final decision I would say that it could be reasonable to send David to the treatment program, however, including high uncertainty.