Statistical Learning Assignment 2

I-Fan Lin s2814412

December 2021

Introduction

In this assignment, the goal is to predict the severity of depressive symptoms after twelve months based on the provided data. In the beginning, the data charastertics will be shortly discussed; secondly, the statistical analysis models will be motivated. Finally, the analysis results will be discussed and evaluated, and then we will use these models to make prediction on a new patient.

Data Description

The provided data contain 1,152 observations, and have 20 predictors, and one response variable. The data is summarized in figure 1. The response variable, dep-sev-fu, meaning the severity of depressive symptoms after twelve months, is non negative integer, and range from 0 to 34; these 20 predictors include ordinal variables, categorical variables and continuous variable.

```
               disType        Sexe          Age             aedu            IDS             BAI              FQ             LCImax       pedigree
 anxiety disorder   :470   female:764   Min.   :18.00   Min.   : 5.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000   No :179
 comorbid disorder  :429   male  :388   1st Qu.:32.00   1st Qu.:10.00   1st Qu.: 8.00   1st Qu.: 5.00   1st Qu.: 2.000   1st Qu.:0.2075   Yes:973
 depressive disorder:253                Median :43.00   Median :11.00   Median :12.00   Median : 9.00   Median : 6.000   Median :0.4259
                                        Mean   :42.11   Mean   :11.76   Mean   :11.62   Mean   :10.45   Mean   : 8.156   Mean   :0.5031
                                        3rd Qu.:52.00   3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:12.000   3rd Qu.:0.8727
                                        Max.   :65.00   Max.   :18.00   Max.   :25.00   Max.   :42.00   Max.   :40.000   Max.   :1.0000
                                        alcohol                    bTypeDep       bSocPhob         bGAD          bPanic          bAgo
 Diagnose alcohol dependent or alcohol abuse:355   Dysthymia             : 19   Negative:716   Negative:852   Negative:750   Negative:1032
 No positive alcohol diagnose              :797   First onset MDD       :345   Positive:436   Positive:300   Positive:402   Positive: 120
                                                  No depressive disorder:390
                                                  Recurrent MDD         :398


      AO          RemDis                                sample      ADuse     PsychTreat    dep_sev_fu
 Min.   : 4.00   FALSE:817   General population            :100   FALSE:698   FALSE:590   Min.   : 0.00
 1st Qu.:12.00   TRUE :335   Primary care                 :508   TRUE :454   TRUE :562   1st Qu.:13.00
 Median :18.00               Specialised mental health care:544                           Median :17.00
 Mean   :20.98                                                                            Mean   :16.91
 3rd Qu.:28.00                                                                            3rd Qu.:20.00
 Max.   :63.00                                                                            Max.   :34.00
```

Figure 1: *Summary Of Data.*

In the following tasks, the data will be split to 1,000 observations for train data, which will be used for training model, and 152 observations for test data, which will be used for comparison of different models.

Exercise 1

Before selecting three methods, I fitted a linear regression with these 20 predictors without interactions. Figure 2 shows the residual plot, and it can be seen that a non-linearity pattern between fitted values and residuals when fitted value is larger than 21. This implies that either the adjustment of the model is needed, or more complicated models can be considered. In this assignment, I am going to select more complicated models.

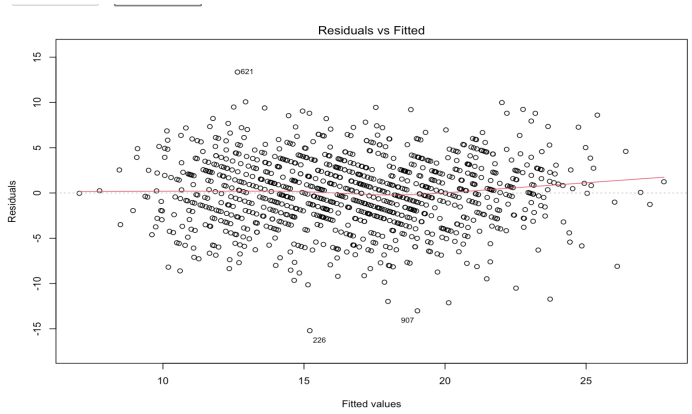


Figure 2: *The Residual Plot of Multiple linear regression.*

I selected random forest, gradient boosting regression, and neural networks for analyzing the data.These methods have an common advantage of modeling non-linearity, which is able to catch the non-linear pattern of the relation in the data. The individual reasons for choosing these methods are discussed as follows:

Random Forests

The idea of the random forests (RFs) in regression problem is to build a lot of independent trees, and then average prediction of these trees as its prediction. Specifically, different bootstrapped samples are used for each tree, and different subsets of the predictors are used to split the tree. The main reason for selecting this method is that it is an ensemble method, whose performance will be more stable compared to only fit a single decision tree. The behavior of creating many trees and average them can effectively reduce the variance, and improve the prediction result, which is our goal in this assignment. Additionally, random forest is flexible in terms of model building, therefore, only some hyperparmeters needed to be tuned. Finally, the predictor selections are not hyperparameters (only need to decide the number of subset of variables being used in each tree), and it can be automatically done by the model.

Stochastic Gradient Boosting

The idea of the gradient boosting regression model (GBM) is like RFs, which using a collection of trees to make prediction. However, unlike RFs, GBM builds tree sequentially, and each tree can learn experience from previous trees because during the training process, the residuals improved (the difference between observed and predicted valued in training data) at t_{th} tree will be the starting point, and then be further improved by $(t+1)_{th}$ tree , and so on. The "stochastic" means the different sub-sample will be randomly selected to build each trees.

The advantage is that the GBM will account for all the predictions of the trees built in the model when predicting a new observation. Therefore, it can be helpful to avoid overfitted, and improve the performance. Finally, like random forests, the predictors selections are not hyperparameters, and it can be done by the model.

Neural Networks

The Neural Networks (NNs) is a model with multiple hidden layers, and each hidden layer has multiple neurons (functions). Each neuron has parameters of weights and bias (intercept). The advantage of the NNs is that the model complexity is quite high for a large number of parameters, which can approximate complicated true function. Also, **Hrushikesh Mhaskar et al.** has showed that deep layers can approximate the class of compositional functions quickly with exponentially lower number of training parameters and sample complexity in comparison to shallow layer (one hidden layer)(Mhaskar, Liao, & Poggio, 2017).

Considering these advantages above and the training set has size of 1,000, which can be seen enough, the method is selected.

The reason why I did not select the generative additive model is that we have 20 predictors. Although it seems not a large number, when interaction of predictors is included, in the extreme scenario there will be $20 + \binom{20}{2} = 210$ predictors. Additionally, the predictors selection may be a challenging work.

Exercise 2

The derivations of hyperparameters for these models are discussed individually.

Random Forests

The most important hyperprameters for RFs is the number of predictors randomly selected as candidates at each split. The default for regression task in R is $\frac{p}{3}$, which is roughly 7 predictors at each split in our data. I define a search space of the candidate number from 4 to 12, and use the 5-folds cross validation to find the optimal one.

The best combination based on the cross validation is 10 predictors (see figure 3), with minimal root-mean-square error (RMSE) of 4.04.

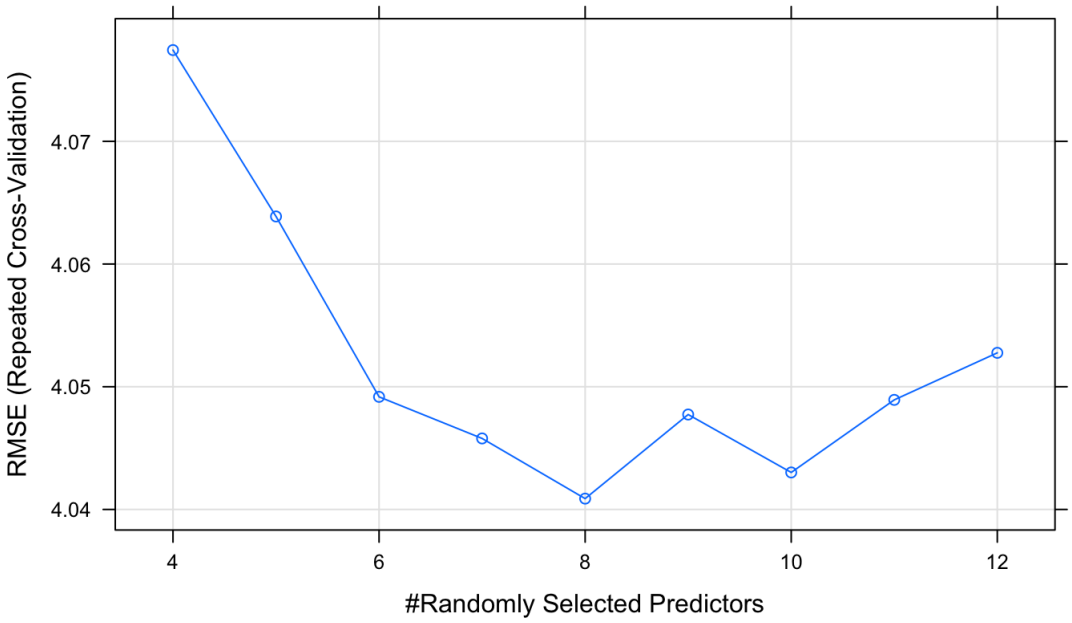


Figure 3: *The Validation Loss By Different Combinations of Hyperparameters.*

Stochastic Gradient Boosting

There are four hyperprameters defined in my search space, the meaning of these hyperparameters are as follow:

1. Shrinkage: Shrinkage is learning rate. Each tree predicts residuals of each training observation, and these residuals will be multiplied with learning rate and be added into the predict value.
2. Number of trees: Number of trees being built in the model.
3. Maximum depth of the trees: When the depth larger than 1, the model will consider two-way interaction effect.

4. Minimum observations in each leaf.

The search space for hyperparameters can be seen in the following table 1. Figure 4 shows validation loss by different combinations of hyperparameters. It can be seen that shrinkage, number of trees, and maximum depth of the trees play more important roles in performance than minimum observations in each leaf did. The best combination based on cross validation is $(shrinakge, ntrees, depth, minObs) = (1000, 2, 0.01, 20)$, with minimal mean-square error (MSE) of 3.94.

Table 1: *The Hyperparameter Search Space for Stochastic Gradient Boosting.*

Shrinkage	$(.1, .01, .005, .001)$
Number of trees	$(10, 100, 1000)$
Maximum depth of the trees	$(1, 2, 3, 4)$
Minimum observations in each leaf	$(10, 20, 30)$

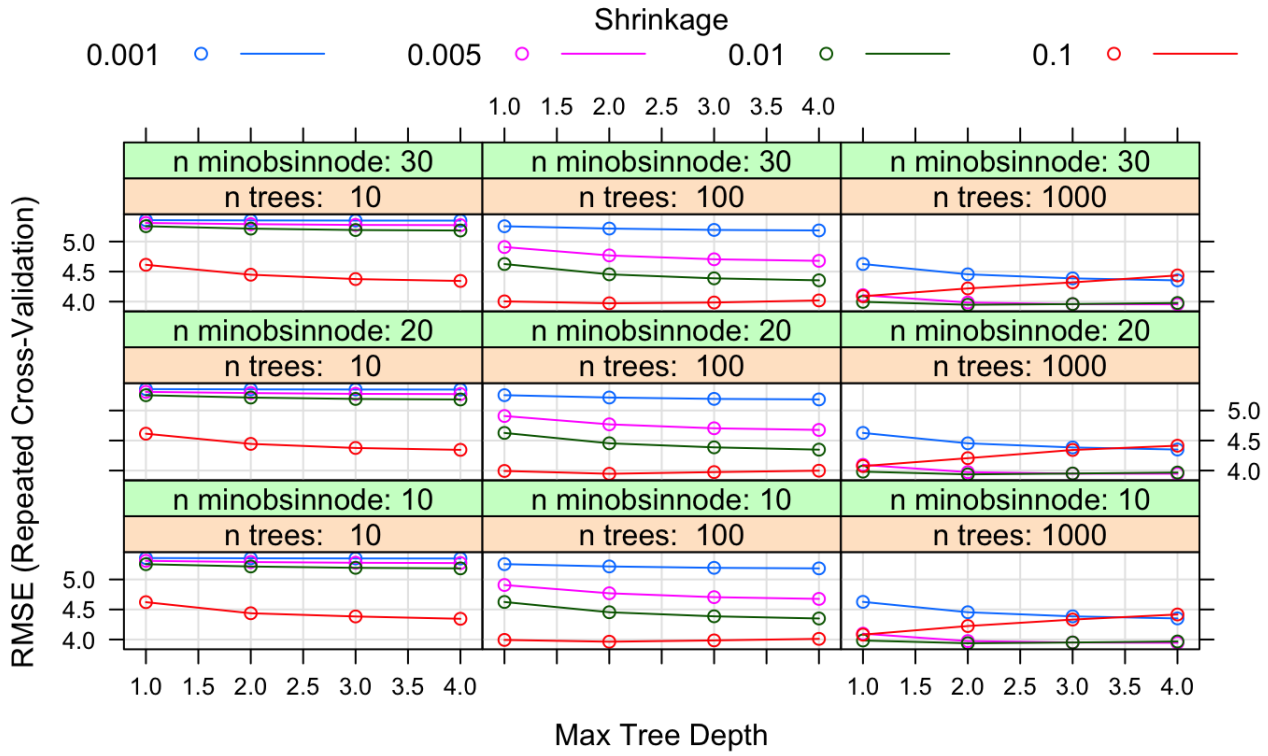


Figure 4: *The Validation Loss By Different Combinations of Hyperparameters.*

Neural Networks

Because there are many hyperparameters for deep neural networks. I only try different architectures of the NNs (see table 2 for architectures and table 3 for common hyperparameters). These models differs in the level of complexity, i.e. the number of parameters. The model 1 has the shallowest layer and the least neurons, and the model 4 has the deepest layer and most neurons.

Table 2: *Neural Networks By Different Hidden Layers Arrangement.*

	Model 1	Model 2	Model 3	Model 4
Hidden layer	(128,64)	(256,128)	(256,128,64).	(256,128,64,32)

Table 3: *The Common Hyperparameters For NNs Models.*

Input layer	
Input shape	27
Hidden layer	
Kernel initializer	Glorot uniform
Activation	sigmoid function
Drop out	None
Output layer:	
Output activation	None
Output shape	1
Optimizer	Adam ($lr = 0.001$)
Loss function	MSE
Maximum epochs*	100

Note: Input shape: the shape is 27 because when we transform the 20 predictors into design/model matrix, the factor variables will become dummy variables, and some of these variables have more than 2 levels such as disType and bTypeDep, which need more variable dimension. Maximum epochs: because early stopping is used, it can be smaller than 100 epochs.

To identify the optimal architecture, I split the training data (1000 observations) in to training set (900 observations) for training and validation set for evaluation, and the optimal one will be the model with minimal MSE in the validation set. Also, the early stopping is applied, so when the validation loss do not improve for certain epochs (10 in my setting), the training will stop, and the parameters generating minimal validation loss will be kept as final parameters for each model, and therefore, the parameters for each model are not necessarily the parameters in 100_{th} epochs. In this task, the best model among these four models is model 1 (see table 4), with MSE of 12.52, and the worst one is model 4 with MSE of 27.55. It shows complex models are not a good choice in the dataset.

Table 4: *MSE in Validation Set.*

	Model 1	Model 2	Model 3	Model 4
Hidden layer	(128,64)	(256,128)	(256,128,64).	(256,128,64,32)
Minimal validation loss (MSE)	12.52	12.76	13.13	27.55

Exercise 3

In the previous exercise, we use the cross validation (for RFs and Stochastic GBM) and validation (for NNs) to find optimal hyperparameters. In this task, these optimal models will be compared based on the test data (152 obervations). Because the response variable dep-sev-fu (the severity of depressive symptoms after twelve months) is non negative integer, the prediction of these models will be rounded to integer. The test MSE can be seen in table 5. The RF performed the best with MSE of 14.17, followed by the Stochastic GBM (15.13), and NNs (16.93). I think the reason the NNs perfomed the worst is that the number of observations is not enough to optimize the parameters, and therefore make it overfitted, leading to NNs approximate the true function relatively poorly.

Table 5: *MSE in Test Set (152 obervations).*

	Random Forests	Stochastic Gradient Boosting	Neural Networks
Test loss (MSE)	14.17	15.13	16.93

To derive the 95% confidence interval(CI) of pairwise differences in predictive performance, the resampling technique is applied. The following process will be repeated 1000 times:

1. Resample 100 observations in test data with replacement.

2. Calculate pair-wise differences of MSE. There are three pairs in total: $(MSE_{RFs} - MSE_{SGBM})$, $(MSE_{RFs} - MSE_{NNs})$, $(MSE_{SGBM} - MSE_{NNs})$ (SGBM is abbrevation of Stochastic GBM).

The 95% CI can be seen in the table 6, the 95% CI of MSE differences for RFs and SGBM includes zero, so we cannot say RFs significantly outperformed than SGBM did, so they both have similar performance. In terms of the other two pairs, the 95% CI of both $MSE_{RFs} - MSE_{NNs}$ and $MSE_{SGBM} - MSE_{NNs}$ are under zero, we can concluded RFs and SGBM significantly outperformed than NNs in the dataset because it implies the RFs MSE and the SGBM MSE is significantly smaller than NNs MSE.

Table 6: *95% Confidence Interval of Pair-wise Differences Of MSE.*

	$MSE_{RFs} - MSE_{SGBM}$	$MSE_{RFs} - MSE_{NNs}$	$MSE_{SGBM} - MSE_{NNs}$
95% CI of MSE difference	(-2.20, 0.35)	(-4.60, -0.91)	(-3.52,-0.14)

Exercise 4 and 5

Because NNs is hard to make interpretations for its deep hidden layers, which involves many additions and compositions of functions, variables of importance and interpretations will be only discussed in the RFs and SGBM in these exercises.

Random Forests

Figure 5 shows the permutation importance (left panel) and improvement in node purity (right panel).

The permutation importance means for a given variable, when the out-of-bags observation are permuted, how much accuracy will decrease. The accuracy is derived by taking average of all trees. The larger the permutation importance, the more important the variables. Based on the measure, the IDS (Testscore on the Inventory of Depressive Symptomatology), and disType (Type of disorder) are the most important predictors to the response variable dep-sev-fu (the severity of depressive symptoms after twelve months).

The improvement in node purity means the total decrease in node impurities from splitting on the variable. The improvement in node purity is derived by taking average of all trees. The larger the value of improvement in node purity, the more important the variables. Based on the measure, the IDS and disType are the most important predictors.

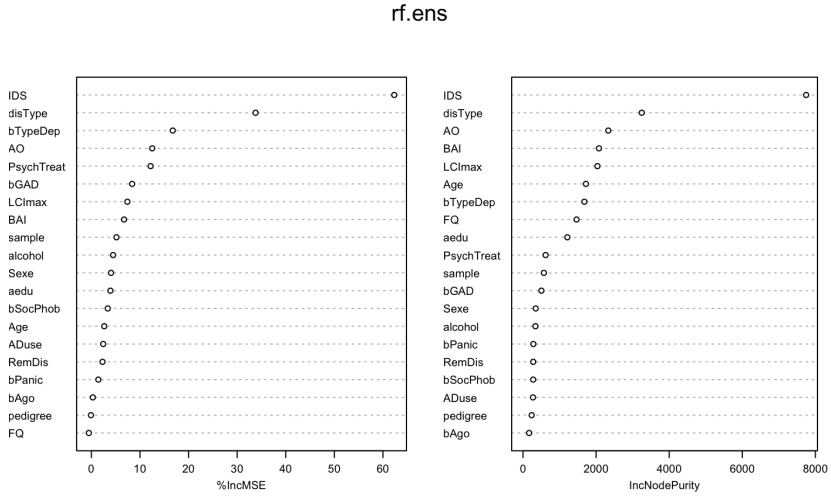
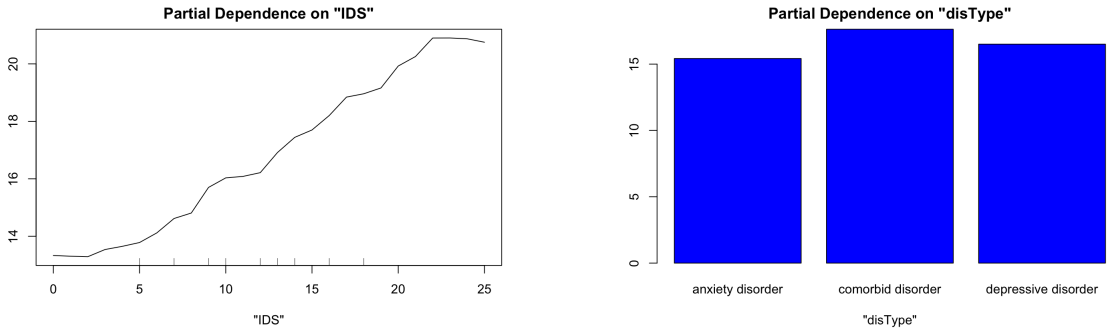


Figure 5: *The importance of variables.*

These two measures align with each other in terms of top two important variables, but other variables have different orders. Figure 6 shows that the effects of these two variables on the dep-sev-fu score. It

shows that IDS tests score has positive linear association with the dep-sev-fu score (left panel) between 0 and 21, and when the IDS score is larger than 21, the increase of IDS score shows no association to dep-sev-fu score. In terms of deType, comorbid disorder has the highest dep-sev-fu score, followed by depressive disorder, and anxiety disorder. However, these illustration assumes that there is without considering interactions of variables, and therefore, if the assumption is violated, the illustration above will be not valid.



(a) *IDS (train data)* **(b)** *disType (test data)*

Figure 6: *Partial Dependency Plot (Random Forests).*

Stochastic Gradient Boosting

The important variables can be observed by using a index called relative influence, which measures the frequency of a variable being used as split over all trees among total nubmer splits of all tress (leaf split is not included). The relative influence of all variables will be summed to 100%. Variables with lager value means they are more important. In the dataset, the IDS, and disType are the two most important predictors (see figure 7)

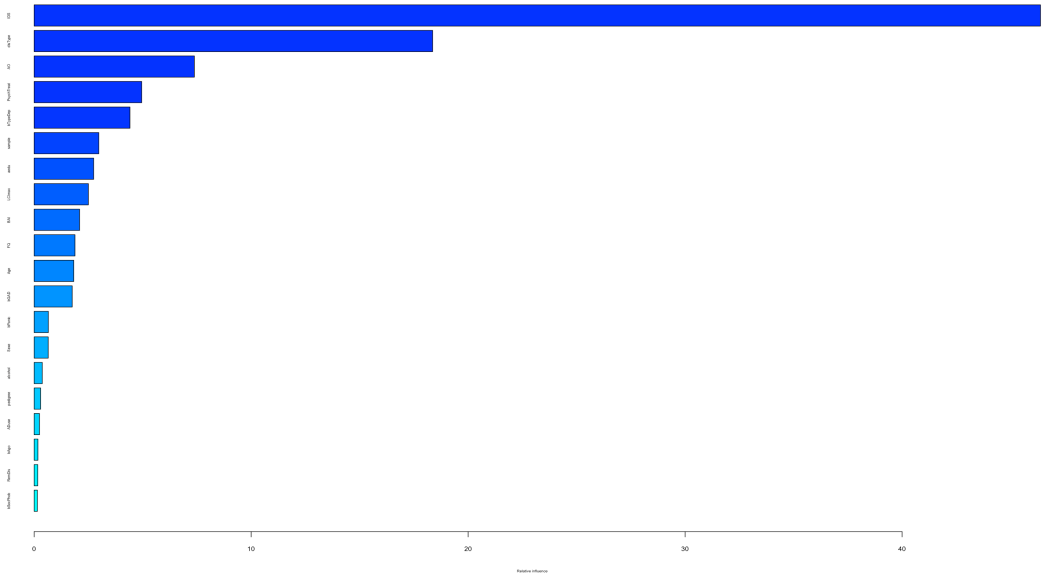
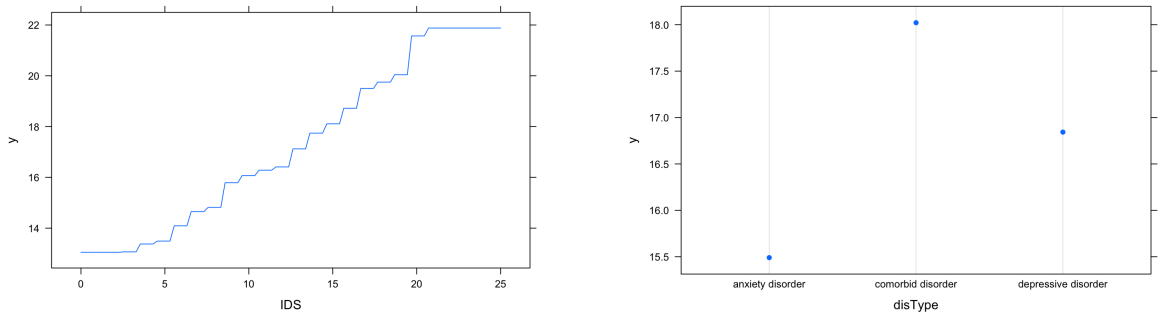


Figure 7: *The Relative Influence Plot of Variables (Unit: %).*

The order from highe to low relative influence is **IDS**, **disType**, AO, PsychTreat, bTypeDep, sample, aedu, LCImax, BAI, FQ, Age, bGAD, bPanic, Sexe, alcohol, pedigree, ADuse, bAgo, RemDis, bSocPhob.

Figure 8 shows that the effects of these two variables on the dep-sev-fu score. It shows that IDS tests score has positive linear association with the dep-sev-fu score (left panel) between 0 and 21, and when the IDS score is larger than 21, the increase of IDS score shows no association to dep-sev-fu score. In terms of deType, comorbid disorder has the highest dep-sev-fu score, followed by depressive disorder, and anxiety disorder. These illustration has the same assumption as previously discussed.



(a) *IDS (train data)*

(b) *disType (test data)*

Figure 8: *Partial Dependency Plot (Stochastic Gradient Boosting Regression).*

Based on the discussion in exercise 4 and 5, I think that all of theses three measures (permutation importance, improvement in node purity, and relative influence) can be taken into consideration of finding the most related predictors because the performance of the RFs and the SGBM are the same in the dataset as discussed in exercise 3. Therefore, if we use the intersection of the most important variables selected by the three measures, then the IDS and disType variables are top two important. Other variables has quite different in orders in these three measures, so it is quite hard to identify whether they are related to response variable.

Exercise 6

Base on the pair-wise prediction performance discussed in exercise 3, the prediction performance of both the RFs and the SGBM are similar, and significantly better than NNs. Therefore, I will only used the two methods to make a decision of the patient, David, by averaging the two predictions and round it, but the prediction of NNs will also be showed as reference.

The predicted value for the new patient is 19, which is derived from averaging the predictions of RFs and SGBM. Because it is larger than 17, it is highly recommendable to refer David to the intensive treatment program.

Table 7: *Predicted Value On New data.*

	RFs	SGBM	Avg. of RFs and SGBM	NNs
prediction	18	19	19*	20

Note: 19 is derived by rounding $18.5 = \frac{18+19}{2}$.

References

Mhaskar, H., Liao, Q., & Poggio, T. (2017). When and why are deep networks better than shallow ones? In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).