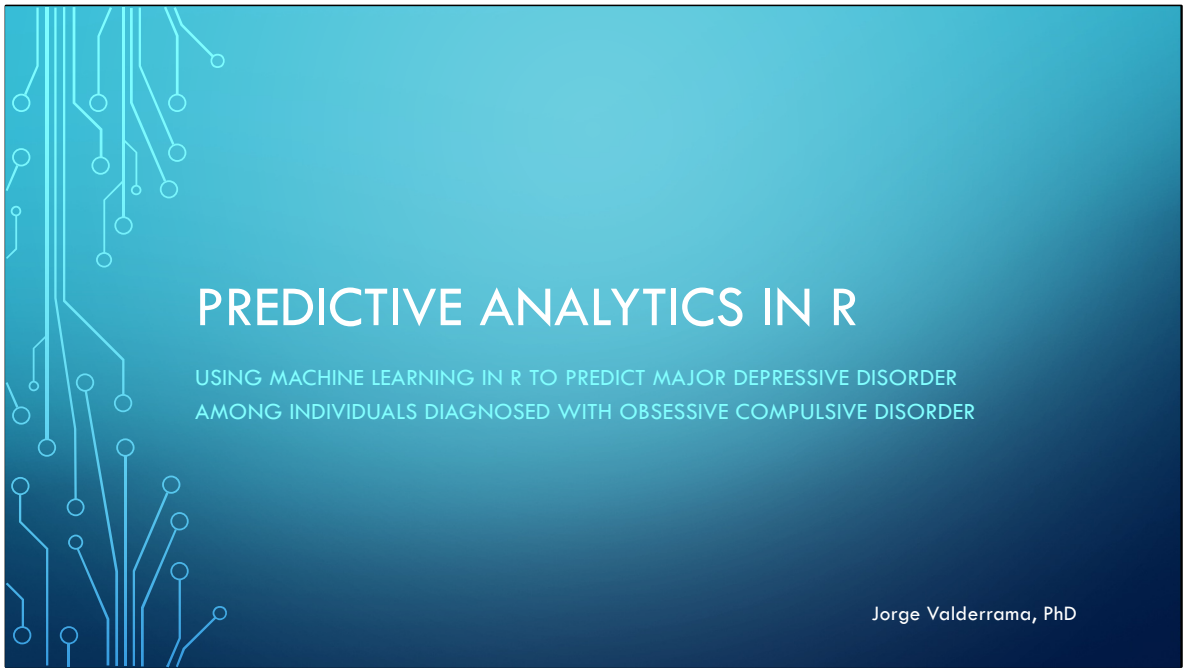

PREDICTIVE ANALYTICS IN R
USING MACHINE LEARNING IN R TO PREDICT MAJOR DEPRESSIVE DISORDER AMONG INDIVIDUALS DIAGNOSED WITH OBSESSIVE COMPULSIVE DISORDER
Jorge Valderrama, PhD

OBJECTIVES

Comprehensive Data Exploration:

- Investigate a dataset comprising individuals diagnosed with Obsessive-Compulsive Disorder (OCD) to enhance our understanding of the factors predicting co-morbid Major Depressive Disorder (MDD).

Advanced Predictive Modeling:

- Employ Machine Learning algorithms in R to predict the occurrence of MDD among individuals with OCD, aiming for a nuanced understanding of the predictive factors.

OBJECTIVES (CONT.)

Algorithmic Evaluation:

- Implement and assess the efficacy of advanced algorithms, including Random Forest, Neural Network, and Xgboost, to predict MDD. Evaluate and compare their accuracy against the model training dataset to ensure robust predictive performance

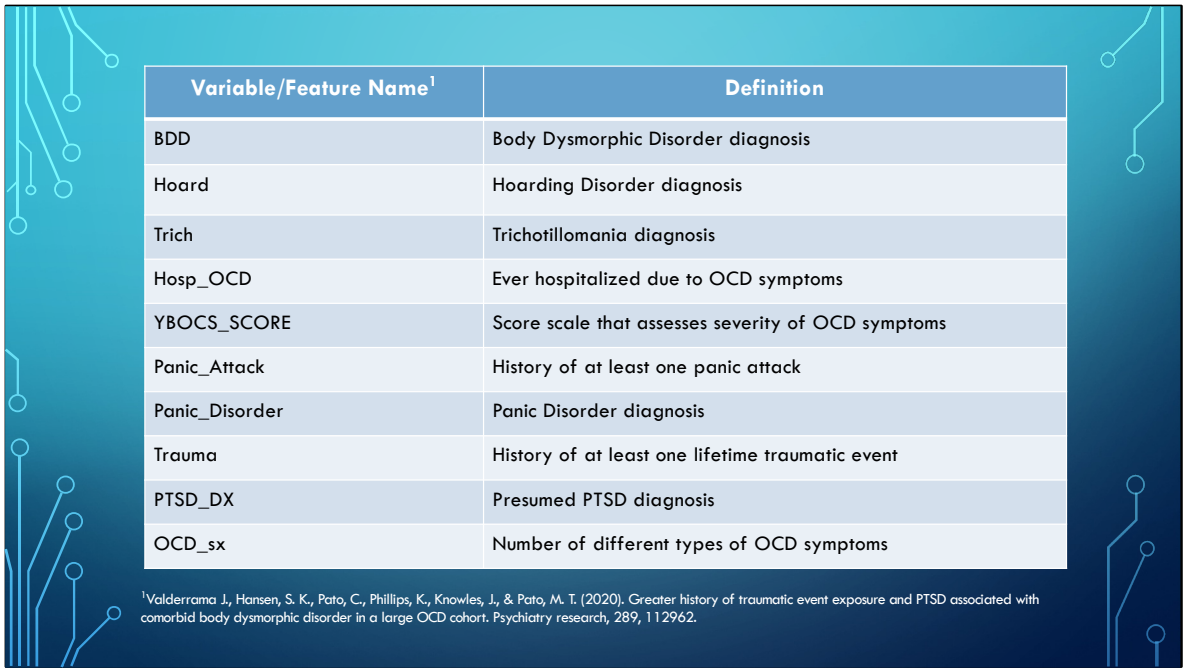
Model Validation with Test Datasets:

- Rigorously validate the predictive models by comparing their performance against independent test datasets to ensure the reliability and generalizability of the developed models.

DATA

- The data includes a subset of 509 individuals with OCD who were recruited as part of the larger Genomic Psychiatry Cohort OCD research study (GPC-OCD). [1]

[1]Valderrama J., Hansen, S. K., Pato, C., Phillips, K., Knowles, J., & Pato, M. T. (2020). Greater history of traumatic event exposure and PTSD associated with comorbid body dysmorphic disorder in a large OCD cohort. Psychiatry research, 289, 112962.



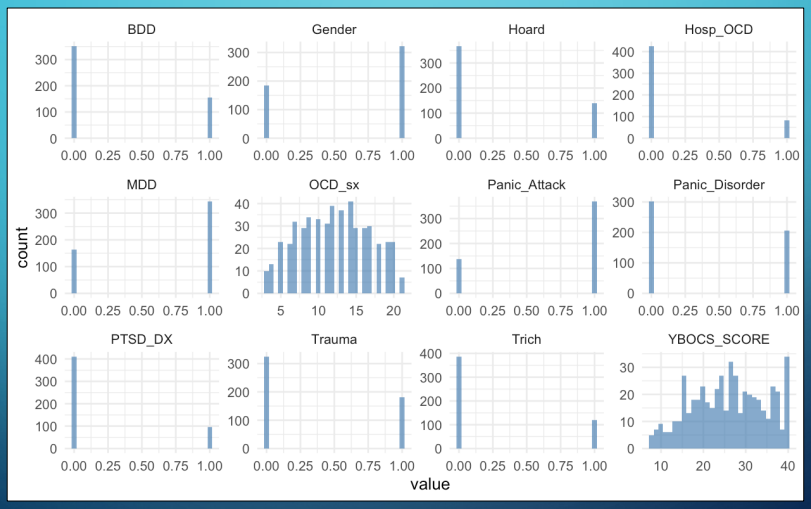
Variable/Feature Name[1]	Definition
BDD	Body Dysmorphic Disorder diagnosis
Hoard	Hoarding Disorder diagnosis
Trich	Trichotillomania diagnosis
Hosp_OCD	Ever hospitalized due to OCD symptoms
YBOCS_SCORE	Score scale that assesses severity of OCD symptoms
Panic_Attack	History of at least one panic attack
Panic_Disorder	Panic Disorder diagnosis
Trauma	History of at least one lifetime traumatic event
PTSD_DX	Presumed PTSD diagnosis
OCD_sx	Number of different types of OCD symptoms

[1]Valderrama J., Hansen, S. K., Pato, C., Phillips, K., Knowles, J., & Pato, M. T. (2020). Greater history of traumatic event exposure and PTSD associated with comorbid body dysmorphic disorder in a large OCD cohort. Psychiatry research, 289, 112962.

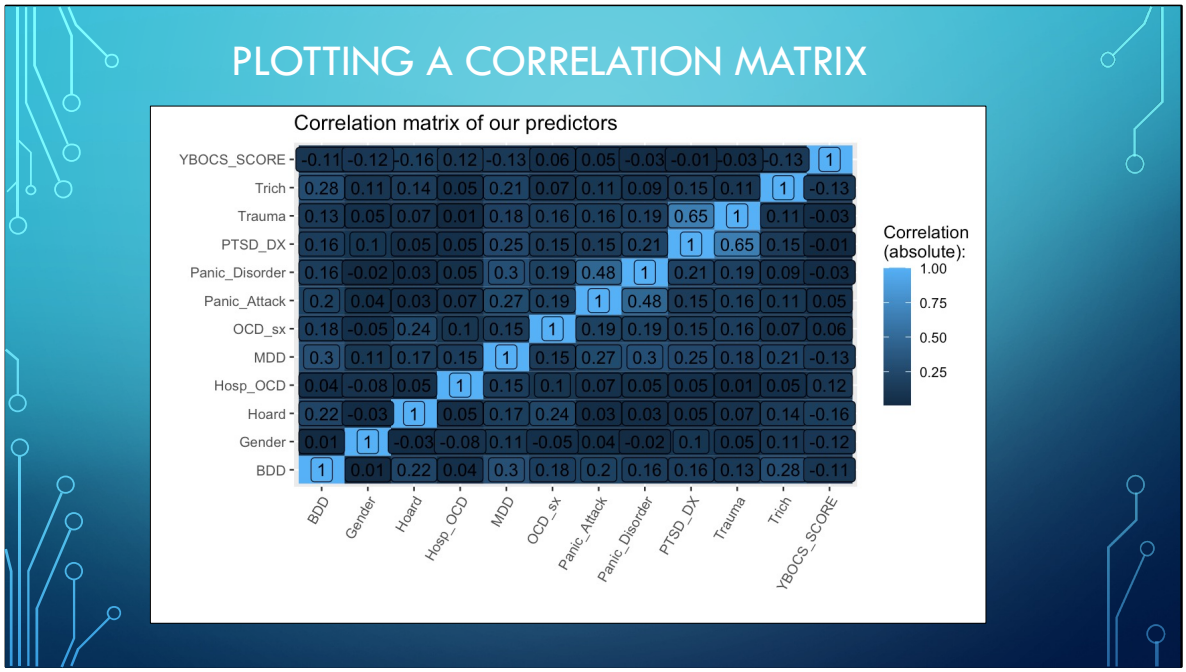
Along with age and gender, here are the list of variables/features in our data set.

Further information about the definition of these variables can be found in the citation at the bottom of the screen.

VISUAL INSPECTION OF FEATURES IN DATASET

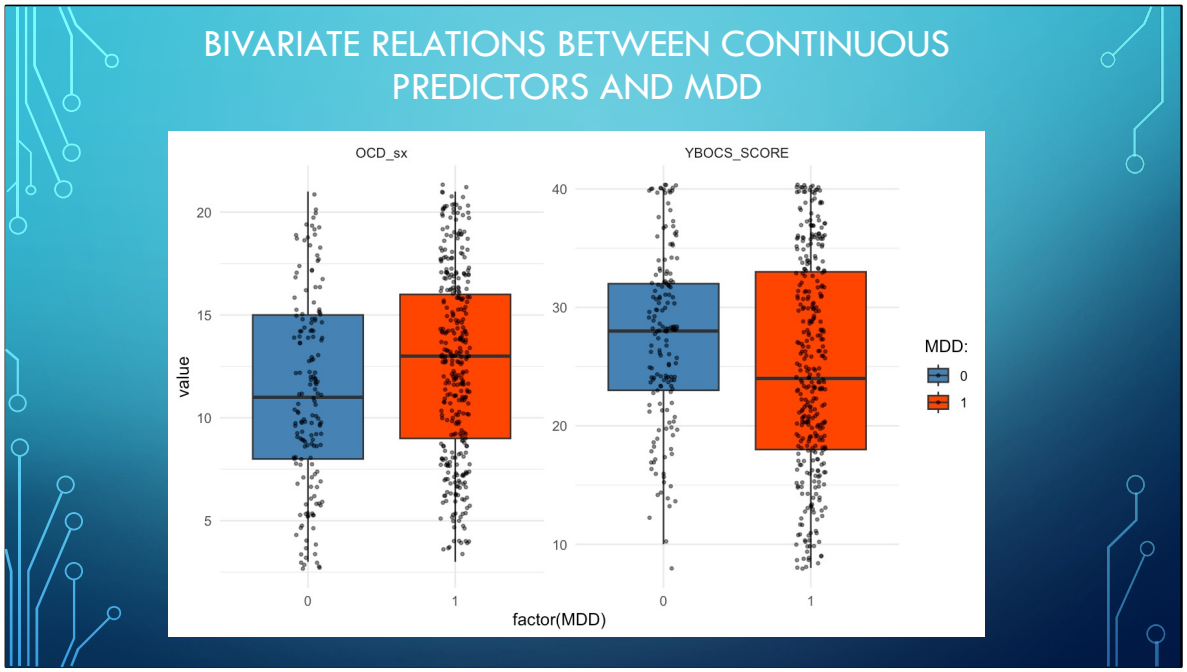


It's important to visualize the features (e.g., variables) in a dataset before building a machine learning model. I first created a plot of histograms. The most important variable to look at is our outcome variable (e.g., our Y), labeled "MDD". You can see that there are more individuals with MDD (n = 300+) than without MDD (n <200). You can also see imbalance in our other categorical variables that have a Yes (1) or No (0) outcome (e.g., BDD, Gender (more females than males), Hoarding, ever hospitalized with OCD symptoms, etc.). With regards to our continuous variables, number of different types of OCD symptoms (OCD_sx) is more or less normally distributed. However, YBOCS score exhibits data that might need some form of normalization or discretization.



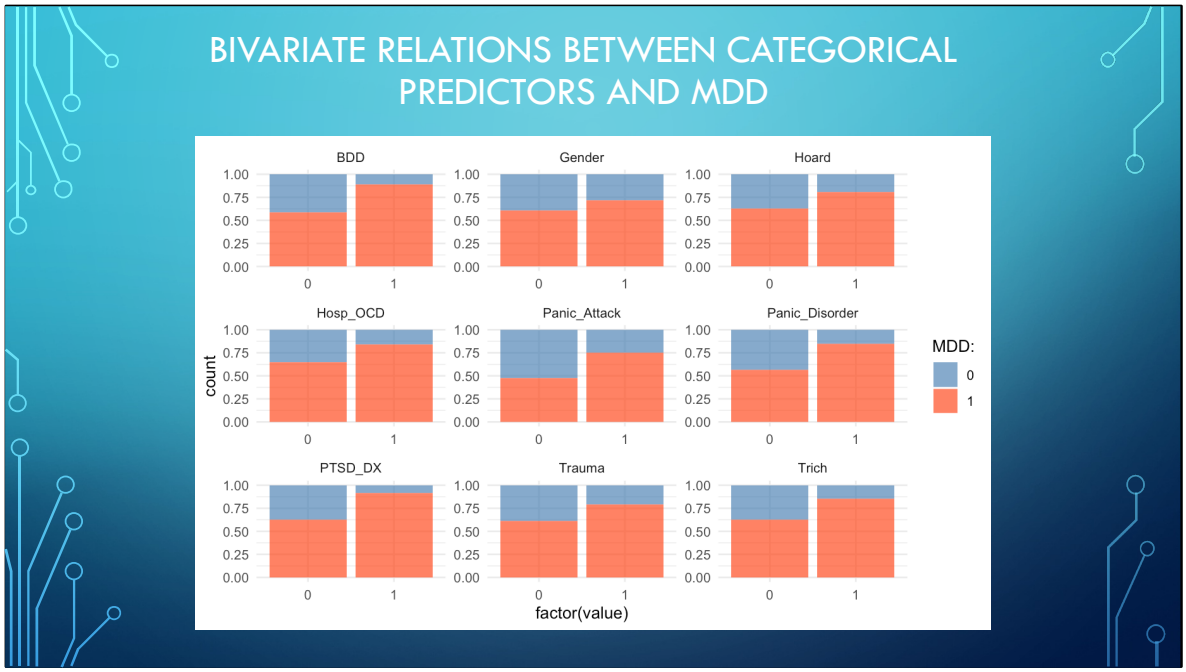
I then plotted a correlation matrix, in order to a) check if there are features that are highly correlated (which is problematic for some algorithms), and b) get a first feeling about which features are correlated with our outcome variable (MDD) and which are not.

We can see that aside from the diagonal (correlation of a variable with itself, which is 1), we have no problematically strong correlations between our predictors (strong meaning greater than 0.8 or 0.9 here).



I then looked at the bivariate relations between the predictors and the outcome. For continuous predictors and a dichotomous outcome (MDD or no MDD), box plots are a good way of visualizing a bivariate association.

You can read this graph as follows: With regards to OCD_sx, participants with MDD (red box) are on average have a higher number of OCD symptoms compared with the participants without MDD (blue box). The thick horizontal line within each box denotes the median. The box encompasses 50% of all cases (i.e. from the 25 percentile to the 75 percentile). The jitter points show you where all of the participants are located within each group. So you see that, yes, participants with MDD patients typically have a higher number of OCD symptoms, but you also have many participants who have lower number of OCD symptoms but have MDD, and of course many participants with high number of OCD symptoms that have no MDD. But comparing the medians, you can see that OCD_sx is a better predictor of MDD than the YBOCS score, which actually shows that people with a high YBOCS score (greater severity of OCD) are less likely to have MDD than to have MDD.



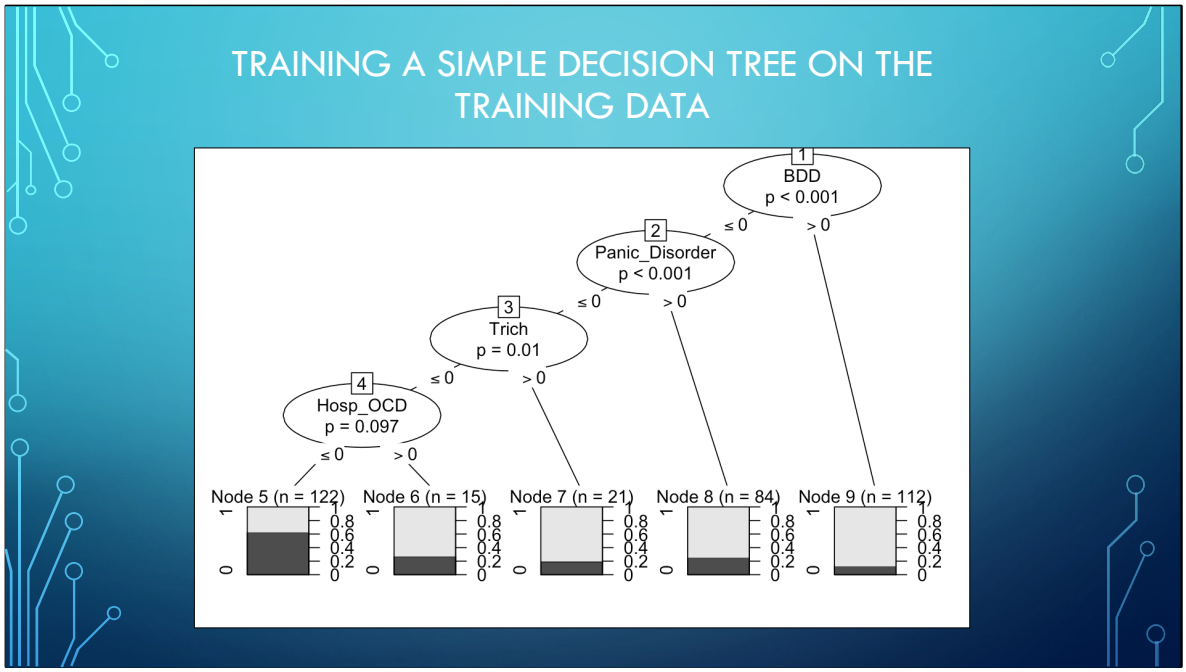
For our categorical variables, I used simple stacked barplots to show the differences between participants with or without MDD. You can see that variables such as BDD, Panic Disorder, PTSD, and trichotillomania are stronger predictors of MDD than Gender or if and individual has ever been hospitalized with OCD.

PARTITION DATA INTO TRAINING AND TEST DATASETS

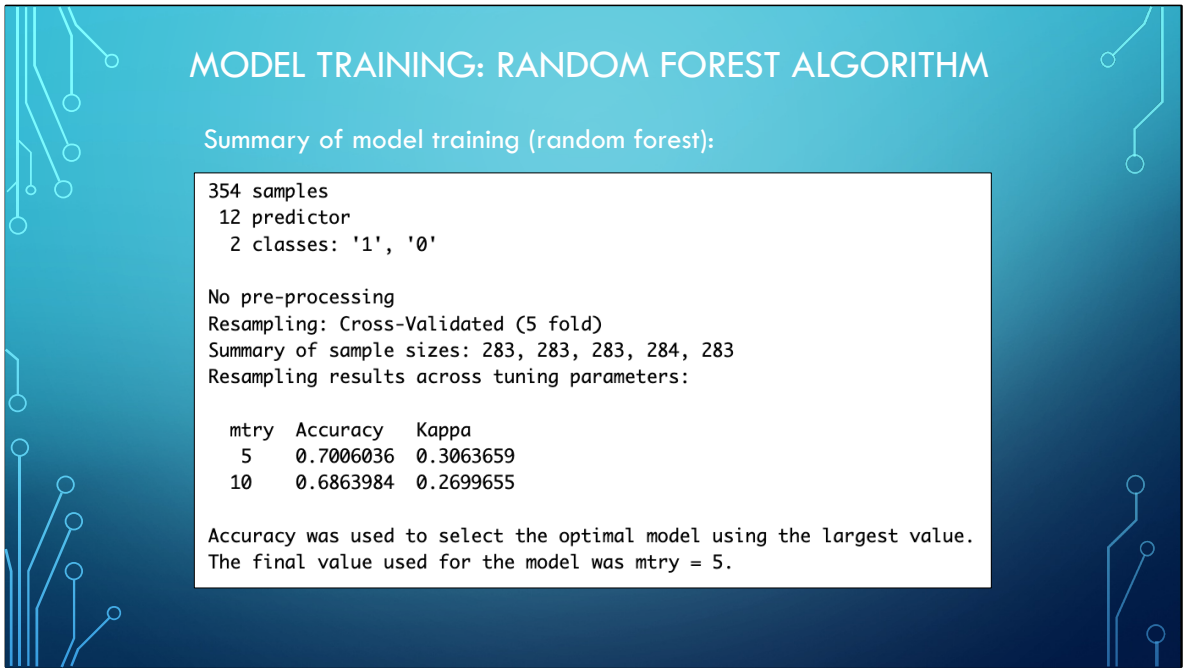
- In R, we want to create training and test datasets to be used for our machine learning algorithms.
- I created a sequence of random numbers which encompass 70% of the dataset, and designated this as “training”, and the rest as a test dataset which will not be touched again until the very end of the analysis.

PRE-PROCESSING

- Since the YBOCS_Score feature was skewed, I created a normalized version of the variable by applying a log transformation
- I kept both variables in the model since the algorithms I used are less sensitive to feature scales (as opposed to a linear regression model)



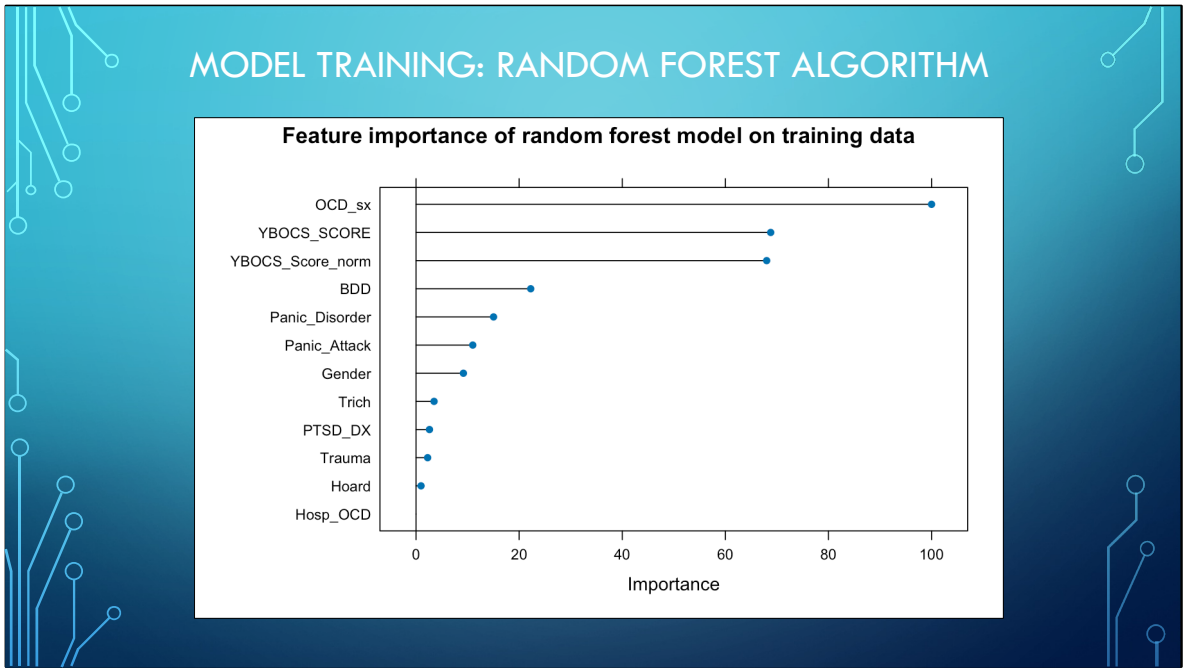
This step is optional but it helped me understand what is going on when I subsequently train a more complex algorithm on the data. Starting from the top, the most important feature that can split the data in two most dissimilar subsets (with regard to how often MDD occurs) is "BDD", i.e. whether a person has a diagnosis of Body Dysmorphic Disorder or not. If a participant has BDD you continue to the right branch of the tree. The bars on the bottom of the tree show the proportion of people who have BDD (light grey) versus those who don't (dark grey). This means that node 9 is the group with the greatest risk for BDD. Whereas the group in Node 5 have the lowest risk. To be in node 5, the person has to have no BDD, no Panic Disorder, no trich, and no history of hospitalization due to OCD symptoms.



A random forest takes a subset of all features (variables) at each tree node. The "mtry" parameter specifies how many of the features to consider at each split. I set this to 5 (it could be anywhere between 1 and 12 (the full number of predictors).

If you look at the accuracy values, you can see that mtry = 5 worked best with the data. On average (of the five runs using cross-validation), 70% of the validation sets classified correctly.

Although this accuracy was obtained with a train/validation split, we still have yet to judge the final evaluation score of the algorithm against the unseen test dataset, because all the patients in the training data were used to train the model at some point, so technically it's not an "out-of-sample" accuracy. But before the final evaluation, I want to try out a two more algorithms.



With a random forest, you can obtain a feature importance plot which tells you which of the variables were most often used as the important splits at the top of the trees.

You can see that, unlike our single decision tree on all of the training data, where "BDD" was the most important feature, it's actually number of different types of OCD symptoms that ends up the most important predictor,. It's important to remember that a machine-learning model tuned for prediction such as a random forest cannot be interpreted as revealing causal associations between the predictors and the outcome. Nevertheless, it can guide clinical practice knowing which features are the most useful for predicting MDD.

MODEL TRAINING: AVERAGED NEURAL NETWORK

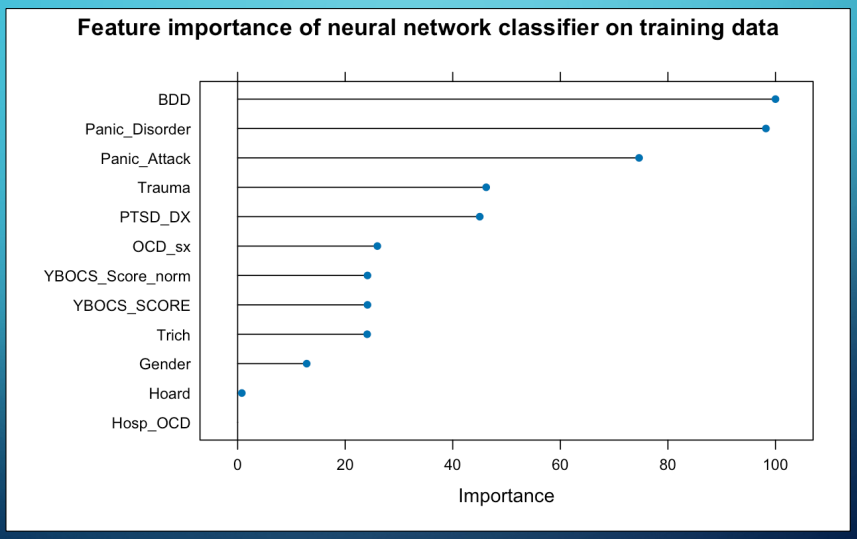
Summary of model training (neural network):

```
Model Averaged Neural Network

354 samples
 12 predictor
  2 classes: '1', '0'

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 283, 283, 283, 284, 283
Resampling results across tuning parameters:

  size  decay  bag    Accuracy   Kappa
   3    0.000  FALSE  0.7005231  0.3175321
   3    0.000   TRUE  0.7006439  0.2903065
   3    0.001  FALSE  0.7032596  0.3432958
   3    0.001   TRUE  0.7315493  0.3835772
   3    0.010  FALSE  0.7173843  0.3598478
   3    0.010   TRUE  0.7118310  0.3452840
   3    0.100  FALSE  0.7286922  0.3781689
   3    0.100   TRUE  0.7343662  0.3748106
   6    0.000  FALSE  0.7030986  0.2978297
   6    0.000   TRUE  0.7033400  0.3059384
```

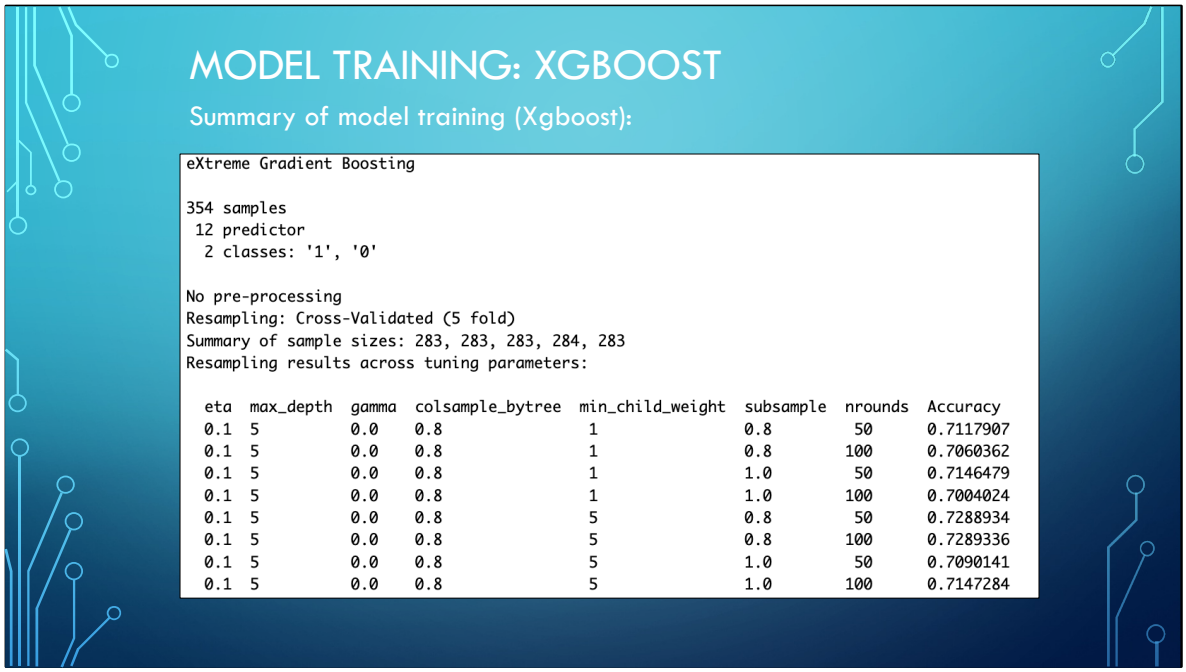
From R output: Accuracy was used to select the optimal model using the largest value. The final values used for the model were size = 3, decay = 0.1 and bag = TRUE.

The best performing model yields a 73.4% accuracy, which is an improvement of over 3% than the random forest algorithm.

MODEL TRAINING: AVERAGED NEURAL NETWORK



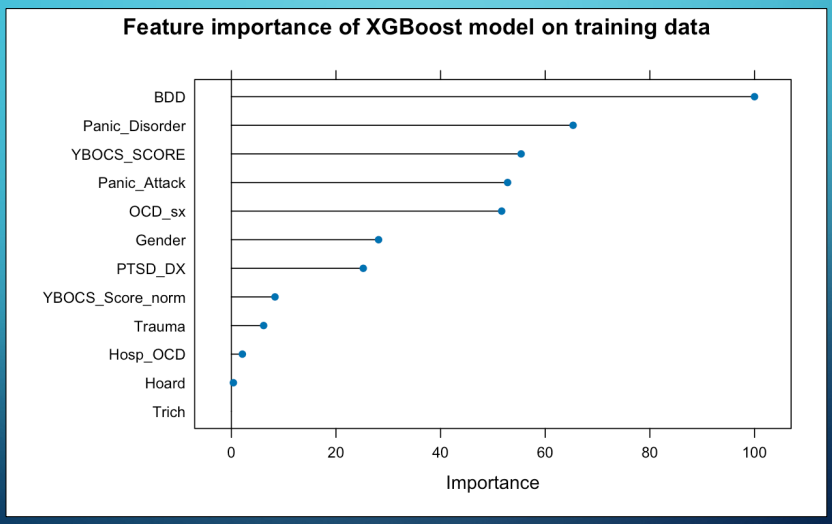
You can see it's a bit different from the random forest feature importance, with BDD, Panic Disorder and a history of at lease one panic attack as the most important features in the model when predicting MDD. Note that with "unstable" methods such as neural networks, if you run the same code 10 times, you can end up with ten (slightly) different feature importance lists, but the general pattern of which features are important and which aren't will be the same.



I tried out one last algorithm. The "(extreme) gradient boosted machines" (xgboost) work similar to a random forest, except they proceed sequentially: A first tree is grown, then more weight is put on the badly predicted samples before the next tree is grown. As a result, in many cases, xgboost outperforms random forests.

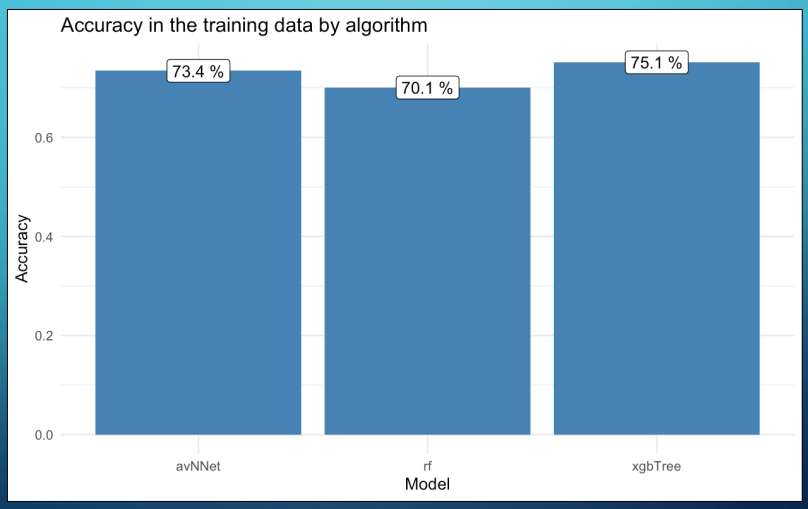
Not shown here: the best performing model yields a 75.1% accuracy, an improvement over the best performing neural network model

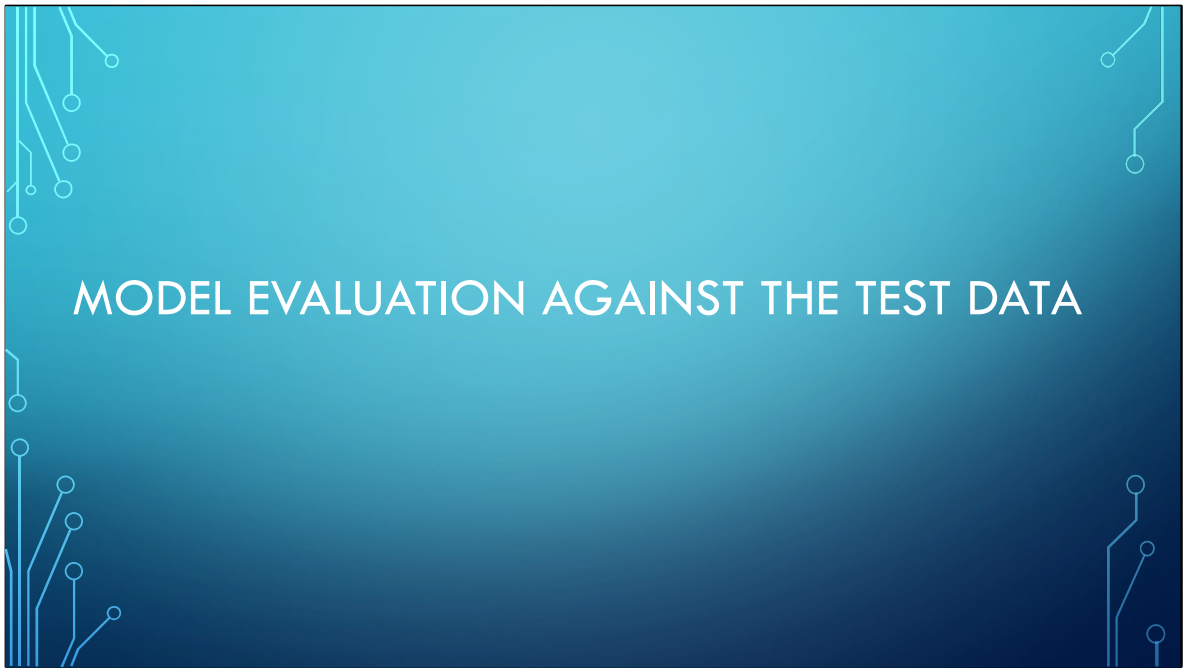
MODEL TRAINING: XGBOOST

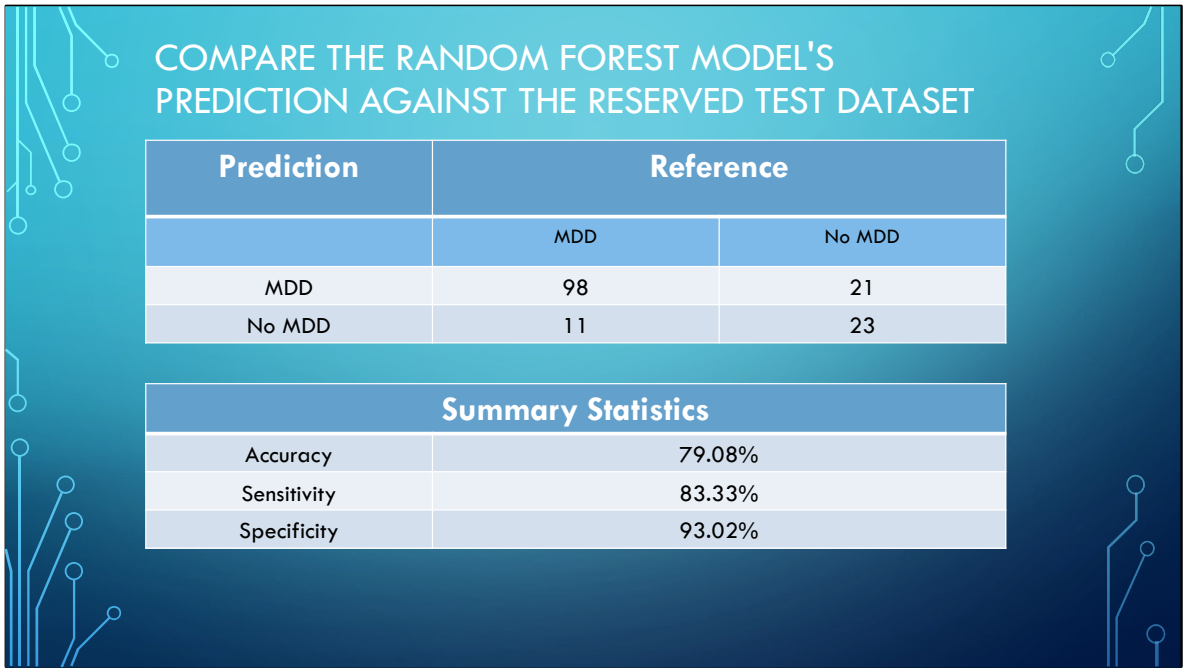


You can see it's a bit different from the random forest feature importance, with BDD, Panic Disorder and a history of at lease one panic attack as the most important features in the model when predicting MDD. Note that with "unstable" methods such as neural networks, if you run the same code 10 times, you can end up with ten (slightly) different feature importance lists, but the general pattern of which features are important and which aren't will be the same.

COMPARING THE PERFORMANCE OF THE THREE ALGORITHMS




MODEL EVALUATION AGAINST THE TEST DATA



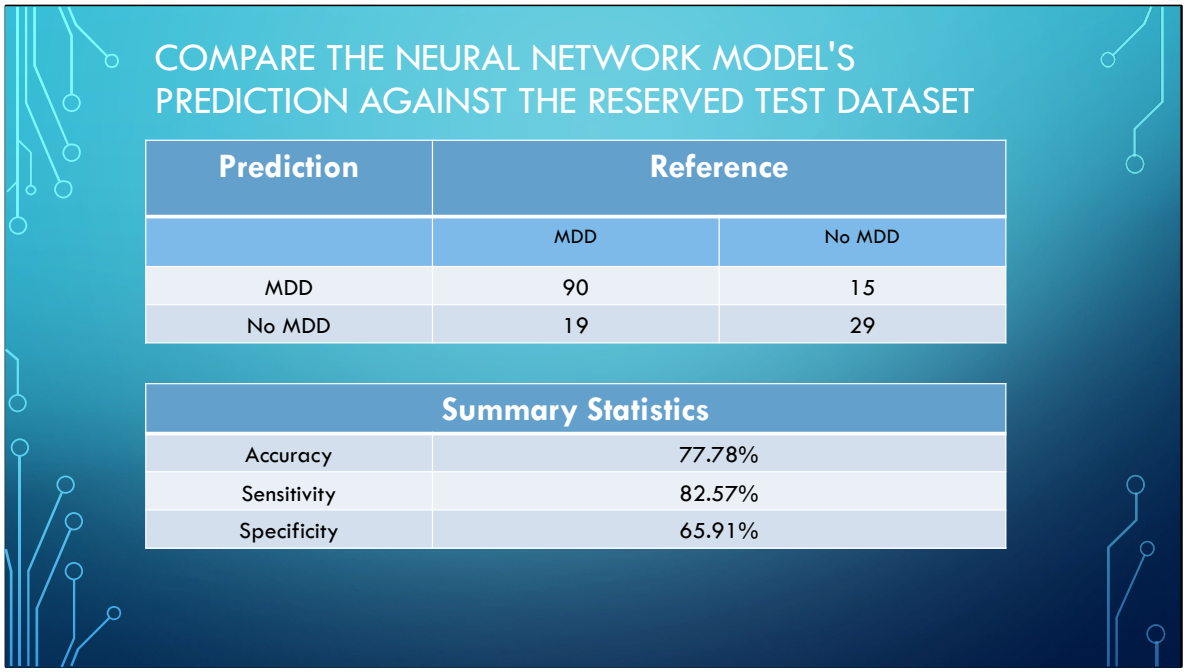
As you can see, our out-of-sample predictive accuracy was 79.08%

The confusion matrix tells us that 98 participants with MDD were correctly classified, and 23 non-MDD participants were also correctly classified, but there were 21 participants where our model thought they had MDD but in reality they didn't, and, conversely, we overlooked MDD in 11 participants.

Sensitivity = how many of the true positive cases are detected, which is a useful indicator if the positive cases are rare, and specificity = how many true negatives are correctly classified.

Other Metrics for the Random Forest Model	
Precision	82.35%
Recall	89.90%
F1 (Harmonic Mean of Precision and Recall)	85.96%

Precision = proportion of true positive predictions relative to all "positive" predictions.
Recall = proportion of true positive predictions relative to all actual positives
F1 = harmonic mean of precision and recall



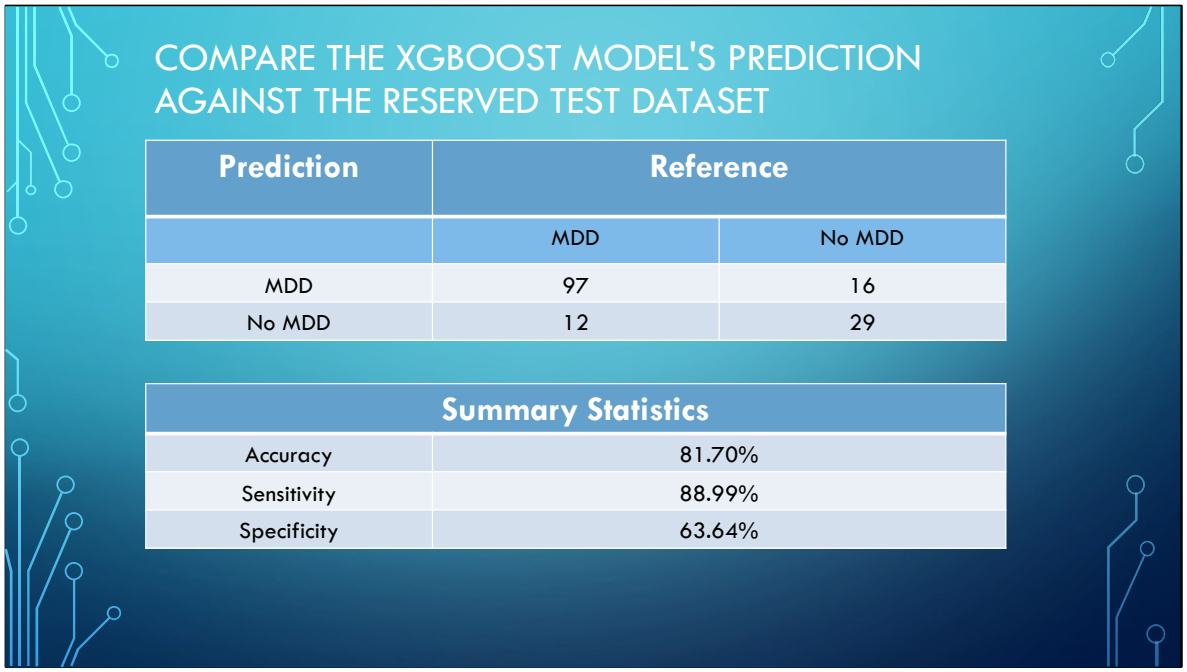
Our out-of-sample predictive accuracy for the neural network model was 77.78%

The confusion matrix tells us that 90 participants with MDD were correctly classified, and 29 non-MDD participants were also correctly classified, but there were 15 participants where our model thought they had MDD but in reality they didn't, and, conversely, we overlooked MDD in 19 participants.

Sensitivity = how many of the true positive cases are detected, which is a useful indicator if the positive cases are rare, and specificity = how many true negatives are correctly classified.

Other Metrics for the Neural Network Model	
Precision	85.71%
Recall	82.56%
F1 (Harmonic Mean of Precision and Recall)	84.11%

Precision = proportion of true positive predictions relative to all "positive" predictions.
Recall = proportion of true positive predictions relative to all actual positives
F1 = harmonic mean of precision and recall



Our out-of-sample predictive accuracy for the neural network model was 81.7%

The confusion matrix tells us that 97 participants with MDD were correctly classified, and 28 non-MDD participants were also correctly classified, but there were 16 participants where our model thought they had MDD but in reality they didn't, and, conversely, we overlooked MDD in 12 participants.

Sensitivity = how many of the true positive cases are detected, which is a useful indicator if the positive cases are rare, and specificity = how many true negatives are correctly classified.

Other Metrics for the Xgboost Model	
Precision	85.84%
Recall	88.99%
F1 (Harmonic Mean of Precision and Recall)	87.38%

Precision = proportion of true positive predictions relative to all "positive" predictions.
Recall = proportion of true positive predictions relative to all actual positives
F1 = harmonic mean of precision and recall

SUMMARY

Strategic Machine Algorithms:

- Developed algorithms designed to offer valuable insights for leveraging predictive analytics in identifying Major Depressive Disorder (MDD) among individuals with Obsessive-Compulsive Disorder (OCD).

Significant Predictive Factor:

- Identified co-morbid Body Dysmorphic Disorder as a pivotal feature in effectively predicting MDD during the model training phase.

SUMMARY (CONT.)

Enhanced Modeling Accuracy:

- Demonstrated superior accuracy of Random Forest and Xgboost algorithms in the evaluation of our model against the test data.

Potential for Real-world Application:

- Recognized the potential for deploying one of these advanced algorithms to create a streamlined screening tool. This tool, based on established risk factors for MDD, could be applied in primary care or community settings, facilitating early intervention strategies.