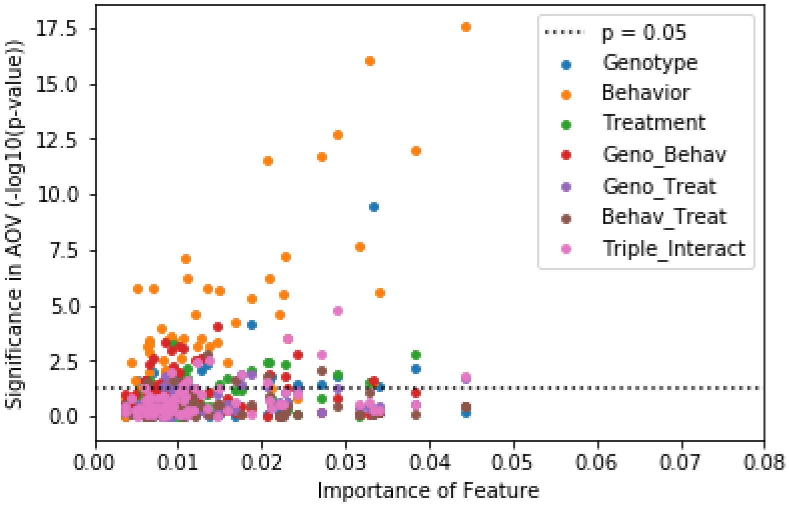
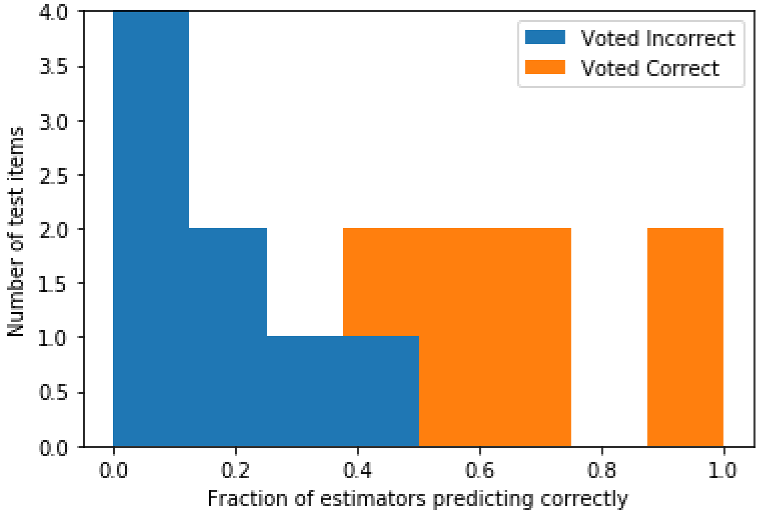
In this dataset, there are 8 groups which are defined by the possible combinations of 3 binary variables: Genotype (control or Down Syndrome model), Behavior (control or learning), and Treatment (saline or memantine). Each of these eight groups contain 7-10 biological replicates (i.e. mice). Ultimately, this is a supervised classification problem.

Initially, I experimented with using a random forest classifier with group as the output. First I did some additional data cleaning, using the median of technical replicates as the value for each biological replicate, rather than the mean I had used for inferential statistics, and then replacing missing values with the median value for each protein. While initially I used the data unscaled, for my final version I used the StandardScaler. I found that the test accuracy (using an 80/20 train/test split) with a single Random Forest classifier was 50-60%. The most important features were enriched for those I had previously identified (by ANOVA) as strongly affected by the Behavior condition.

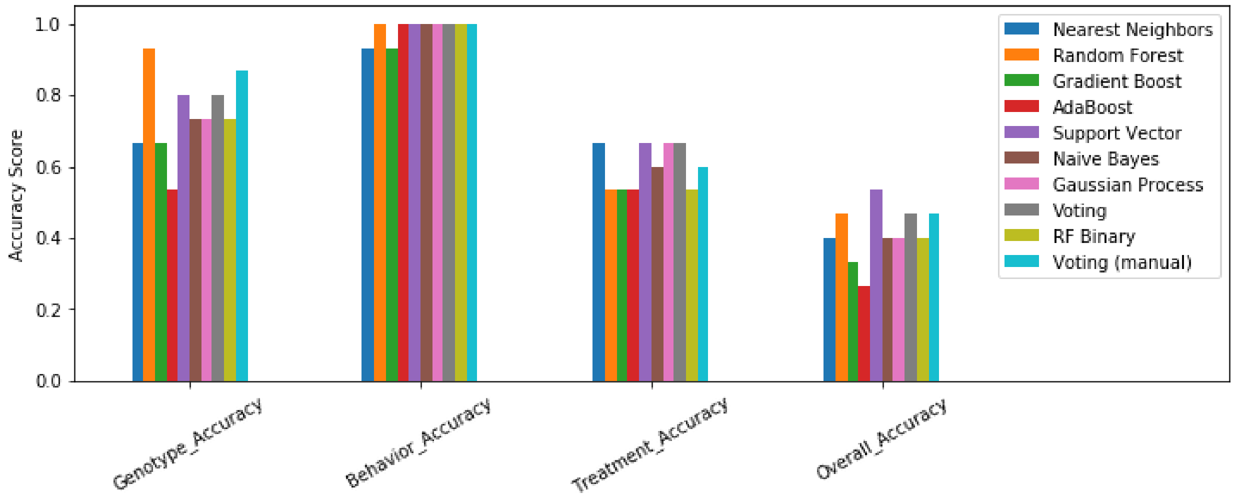


Since the eight groups actually represent a combination of three binary variables, I was curious how well I could predict group membership by independently predicting these three variables. I used Random Forest classifiers with the same hyperparameters as I used for the previous analysis and then combined the results to predict the group. While this method doesn't take into account any interaction between the variables, it increases the sample size per group from 7-10 to 34-38, increasing the ability of the classifier to reliably distinguish between the groups. As a result, it is nearly as good at predicting group membership as the classifier that looks at the final group.

In an attempt to improve the accuracy of my classification, I used a voting classifier to combine many different types of classifier, weighted by their performance in 4-fold cross-validation. I combined Nearest Neighbors, Random Forest, Gradient Boost, AdaBoost, Support Vector, Naive Bayes (Gaussian), and Gaussian Process classifiers. Interestingly, there were not large differences in the test accuracy of any given classifier when run independently, and combining them didn't even boost accuracy, suggesting that the hard-to-classify members may be the same regardless of method. When using a manual vote, I didn't see a clear separation into easy and hard items, though:



When I deconvolved the group predictions into the three variables, all classification methods I tried did well at predicting the Behavior variable and worse at predicting Genotype or Treatment:



Evidently the proteins levels in this dataset overall vary more with learning than according to either the genotype or the treatment. This is consistent with my initial data visualization in which I could see multiple proteins that appeared to vary with learning and did not visually identify those that differed according to the other variables. Likewise, my inferential statistics found that many more proteins were significantly affected by Behavior than by Genotype, Treatment, or the interactions between variables.