

Machine learning for the microbiome!

Machine learning super quick

- Lots of buzzwords (AI, ML, deep learning) that overlap
- Essentially, “learning from data”
- In practical terms, there’s two types we would look at (1 here)
- Supervised- treatments have labels you can learn from, like healthy/unhealthy, etc. (our main focus via SIAMCAT)
- Unsupervised – Clustering to group data into groups
 - Dimension reduction (PCA)

What are some of the benefits of ML?

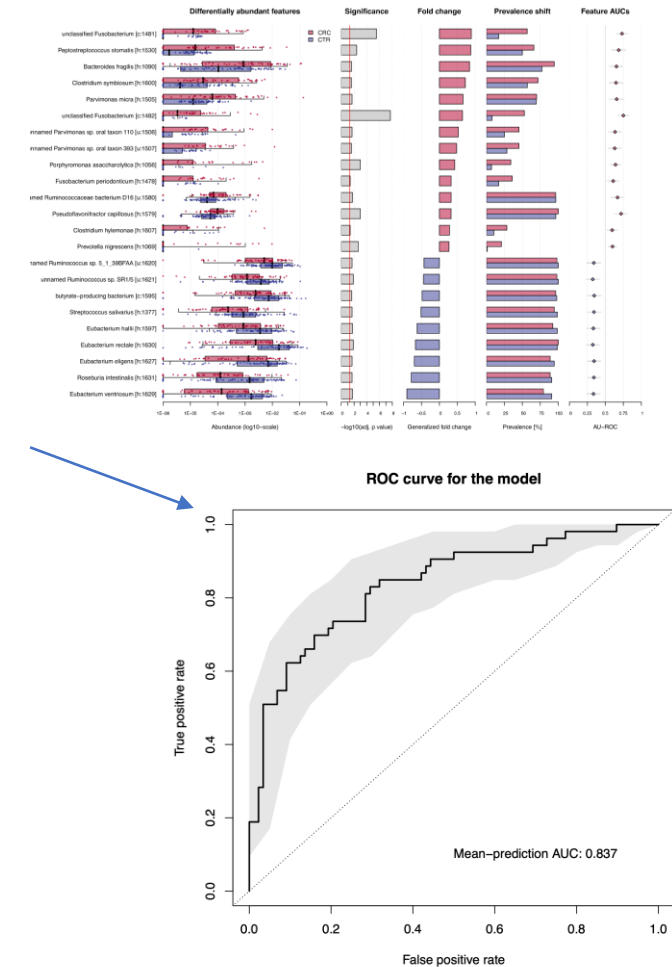
- Feature selection (sometimes!)
 - Some methods provide metrics giving the ranking of the most relevant features (here, ASVs)
- Reuse data via cross validation (CV)
 - CV lets you resample your data to get better estimations (5-10 folds are usually fine)
- Some of the methods are specifically good for high dimensional data
 - Pretty much all the models provided by `siamcat` fit this criteria
 - Lasso, elastic net, ridge regression, random forest (there are others, these are just the ones in `siamcat`)

What can SIAMCAT (potentially) answer?

- Associations between the feature of interest
 - and the ASV/OTUs

Overall prediction
ROC plots-

Upper left is
“perfect
classification”



And the mess-Interpretation plot (ASV X covars)

- Look at clump
- Brown/blue
- In assoc
- le upper
- Right hand

