Team Meeting

29 SEPTEMBER 2022 / 14 :00 / CONFERENCE ROOM

Attendees

Ni, Stefan, Xavier, Ate, Zexi, Jiadong, Shuai

Agenda

- Results for 833
- Results for full datasets
- Queries
- Next step

Results for 833

Available datasets: 5 datasets (E-Risk, BSGS, Denmark, E-MTAB, AMDTSS)

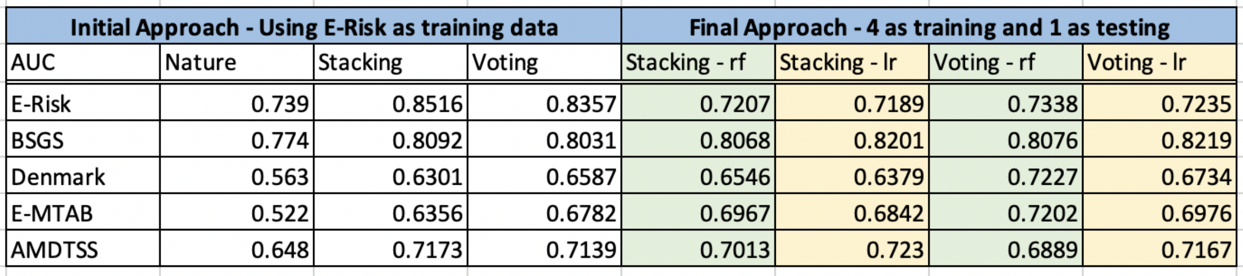
Initial Approach

- Number of variables: 833
- Train data: Erisk

Final Approach

- Variables are selected by rf (random forest) / lr (logistic regression), and the number of variables is reduced from 833 to ~300
- Concating four of them as training and the left one as testing.
  Eg. Training: BSGS, Denmark, E-MTAB, AMDTSS
  Testing: E-Risk

Results: Final Approach is preferred, voting & rf shows a better performance



Initial Approach - Using E-Risk as training data				Final Approach - 4 as training and 1 as testing			
AUC	Nature	Stacking	Voting	Stacking - rf	Stacking - lr	Voting - rf	Voting - lr
E-Risk	0.739	0.8516	0.8357	0.7207	0.7189	0.7338	0.7235
BSGS	0.774	0.8092	0.8031	0.8068	0.8201	0.8076	0.8219
Denmark	0.563	0.6301	0.6587	0.6546	0.6379	0.7227	0.6734
E-MTAB	0.522	0.6356	0.6782	0.6967	0.6842	0.7202	0.6976
AMDTSS	0.648	0.7173	0.7139	0.7013	0.723	0.6889	0.7167

Results for Spartan

HyperParameter Tuning and fit Random Forest(Spartan)
ERISK training auc: 0.60909 development auc: 0.689
E-MTAB training auc: 0.6951 development auc: 0.7081
AMDTMSS: training auc: 0.680749 development auc: 0.703571
BSG and Denmark are still running, taking 2 days to run, 3 or more hours to queue.

Next step

- Select variables using rf, reduce variable size to no more than 1000 variables
- Perform the same thing we did for 833 to the new selected variables and check the auc

Notes

- About the final presentation, do you prefer it to be live or we should share the recording?
-

Action Items

1.