Interpretable machine learning algorithms for understanding factors related to childhood autism

Toby Dylan Hocking
toby.hocking@nau.edu
toby.hocking@r-project.org

February 25, 2024

Motivation and data for predicting childhood autism

- We have data from the National Survey of Children's Health.
- Each year a number of people fill out the survey (rows), and we have data for their responses (columns).
- One column, k2q35a "Autism ASD" (Yes or No) represents if the child has Autism.
- **Data pre-processing**: operations prior to machine learning.
- **Prediction accuracy in a given year**: can we predict this response (output/label/dependent), given the others? (inputs/features/independent)
- **Model interpretation / feature selection**: which inputs are most useful for prediction?
- **Similarity/difference between years**: Can we train on one survey year, and accurately predict on another?

Data pre-processing

Prediction accuracy in a given year

Model interpretation / feature selection

Similarity/difference between years

Discussion and Conclusions

Data pre-processing

survey_year	data.type	nrow	ncol	prop.Autism	prop.NA.rows	prop.NA.cols
2019	raw	29433	445	0.0296	1.0000	0.8989
2019	processed	18202	377	0.0300	0.0000	0.0000
2020	raw	42777	445	0.0298	1.0000	0.8989
2020	processed	27808	373	0.0308	0.0000	0.0000

Data pre-processing

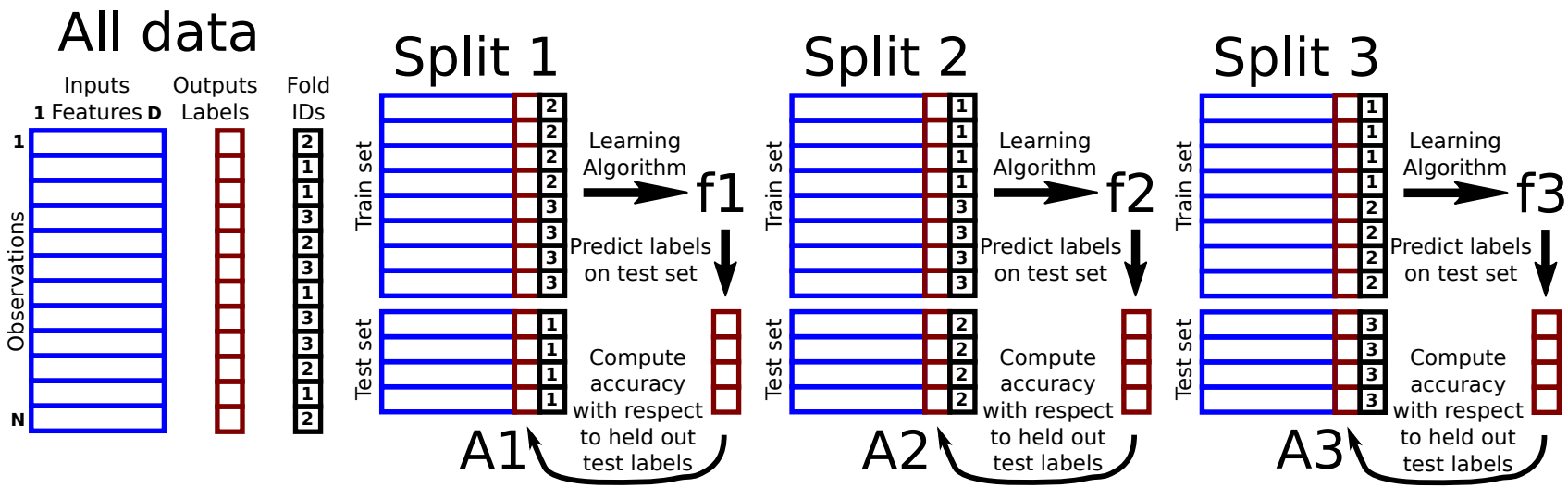
Prediction accuracy in a given year

Model interpretation / feature selection

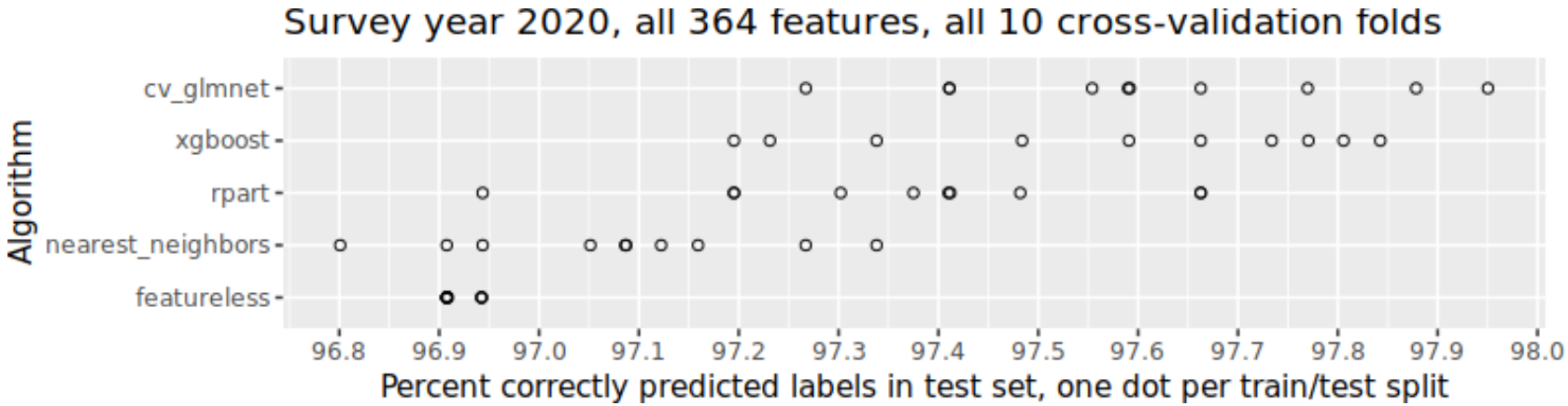
Similarity/difference between years

Discussion and Conclusions

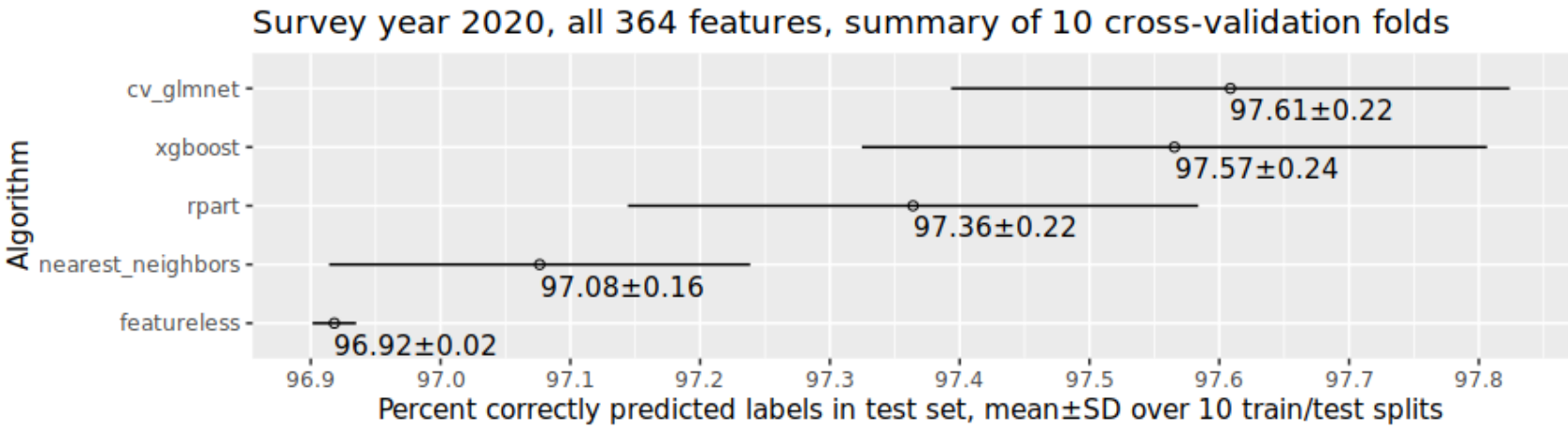
Cross-validation



Cross-validation



Cross-validation

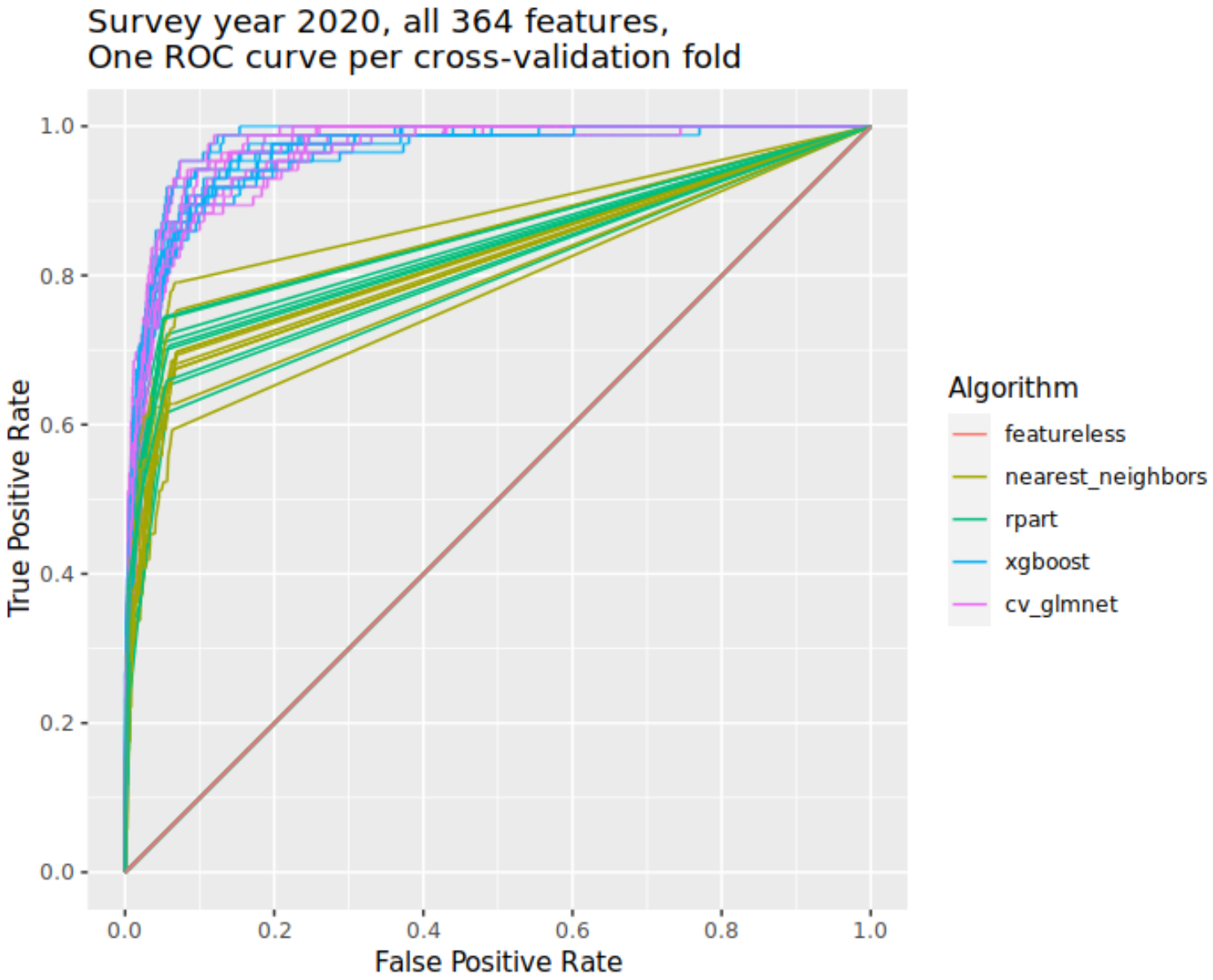


Confusion matrix and error rates

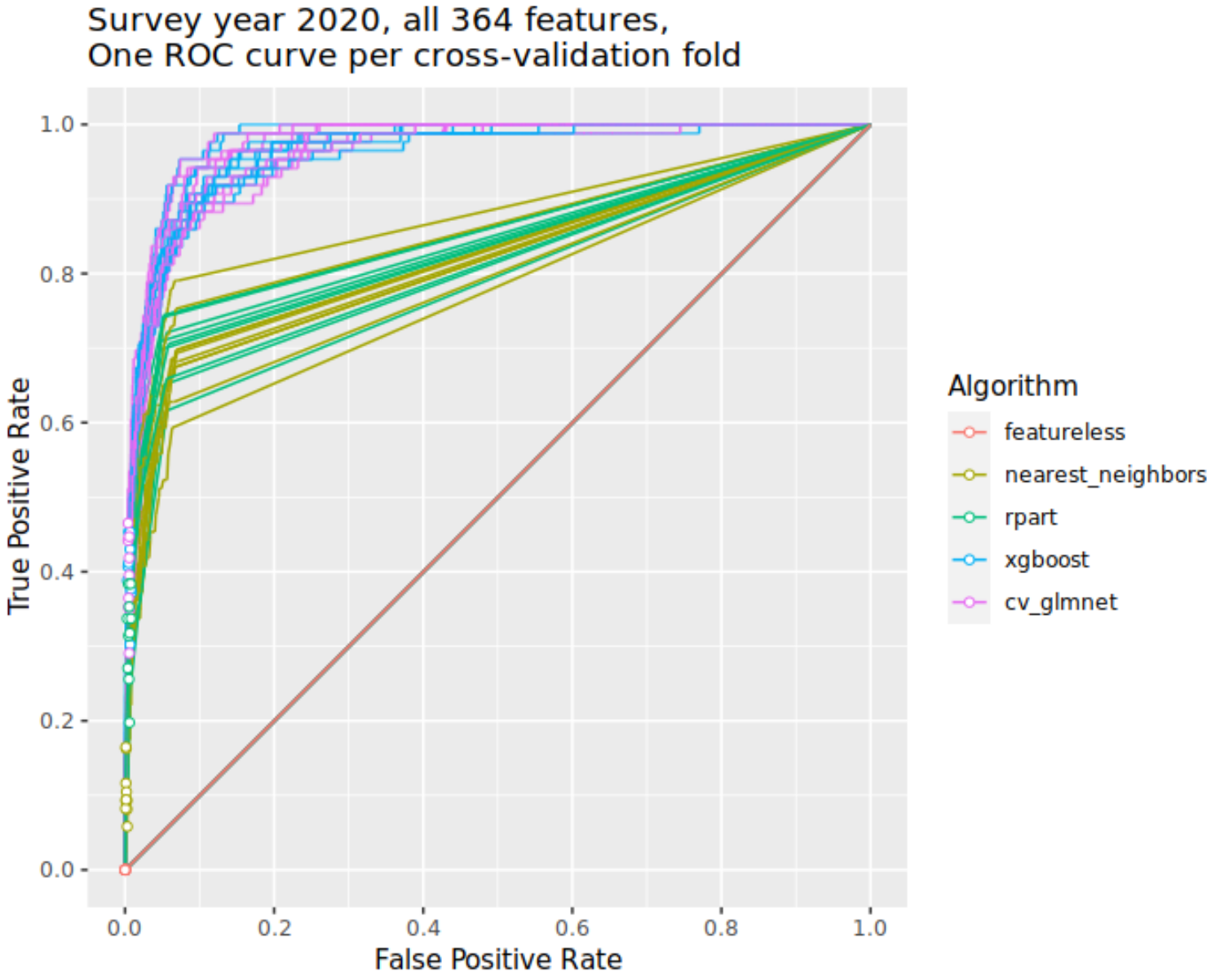
	Label 0	Label 1
Predict 0	True Negative (TN)	False Negative (FN)
Predict 1	False Positive (FP)	True Positive (TP)

- Each has a corresponding rate which is a proportion between zero and one, for example FPR=False Positive Rate.
- Rates are related, TPR=1-FNR quantifies accuracy for positive labels, and TNR=1-FPR is for negative labels.
- TN/TP are good (want to maximize), whereas FP/FN are bad (want to minimize).
- Ideal rates are FPR=0 and TPR=1 but that is not possible to achieve in most real data.
- Receiver Operating Characteristic (ROC) curves trace TPR as a function of FPR, for every threshold of the learned prediction function $f(x)$.

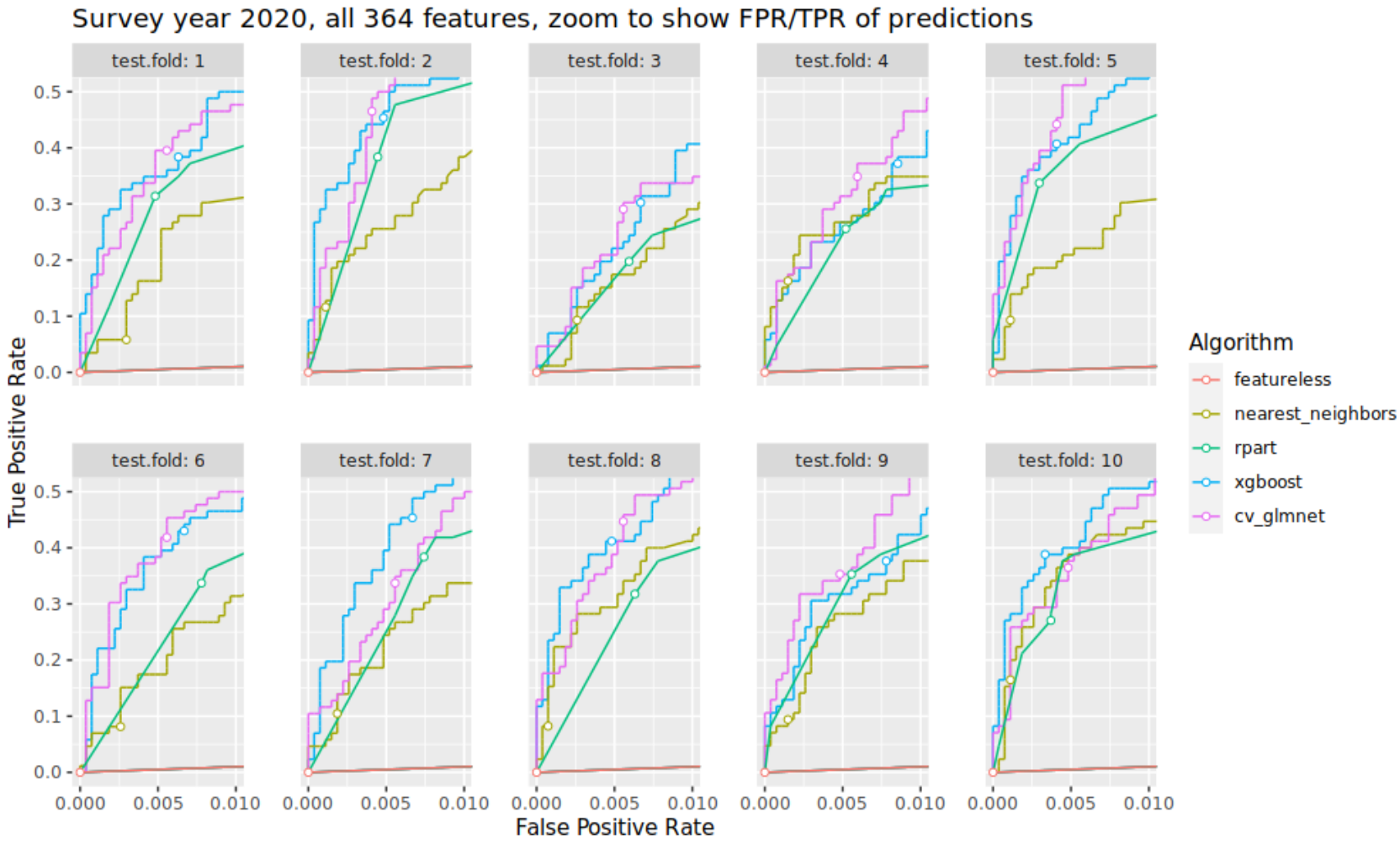
ROC curves show all tradeoffs between TPR and FPR



Default prediction threshold can be viewed as a dot

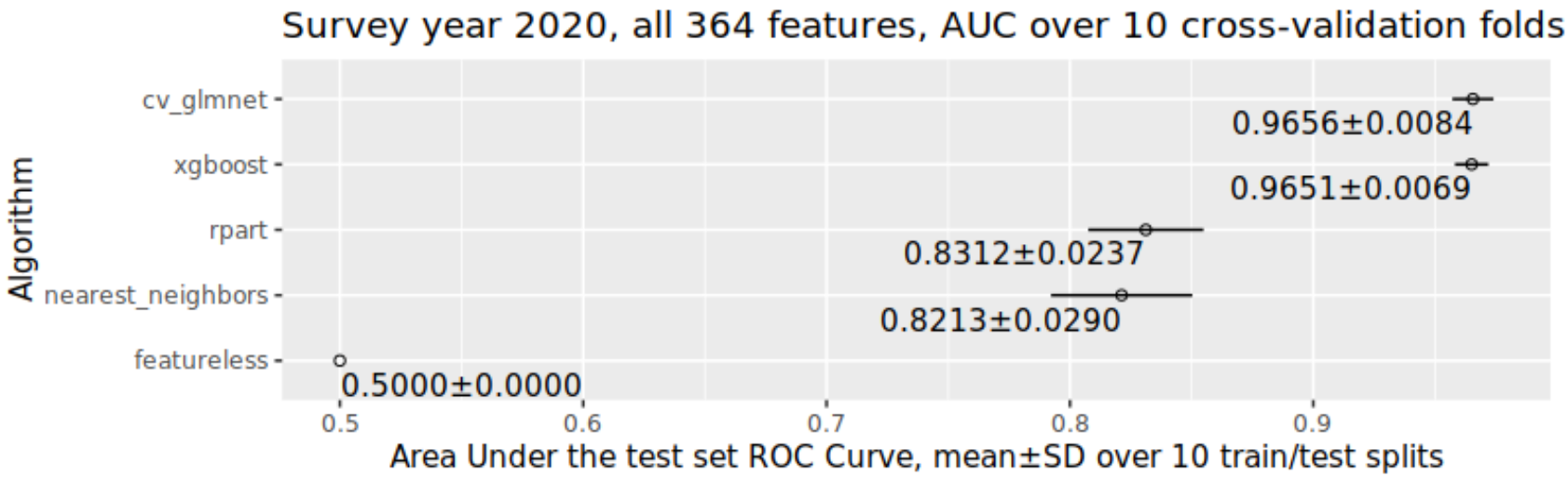


Default prediction threshold can be viewed as a dot



Relatively small FPR because there are so few positive labels (Autism=Yes only 3% of 27808 rows in 2020).

Area Under ROC Curve (AUC) quantifies accuracy over all thresholds



Data pre-processing

Prediction accuracy in a given year

Model interpretation / feature selection

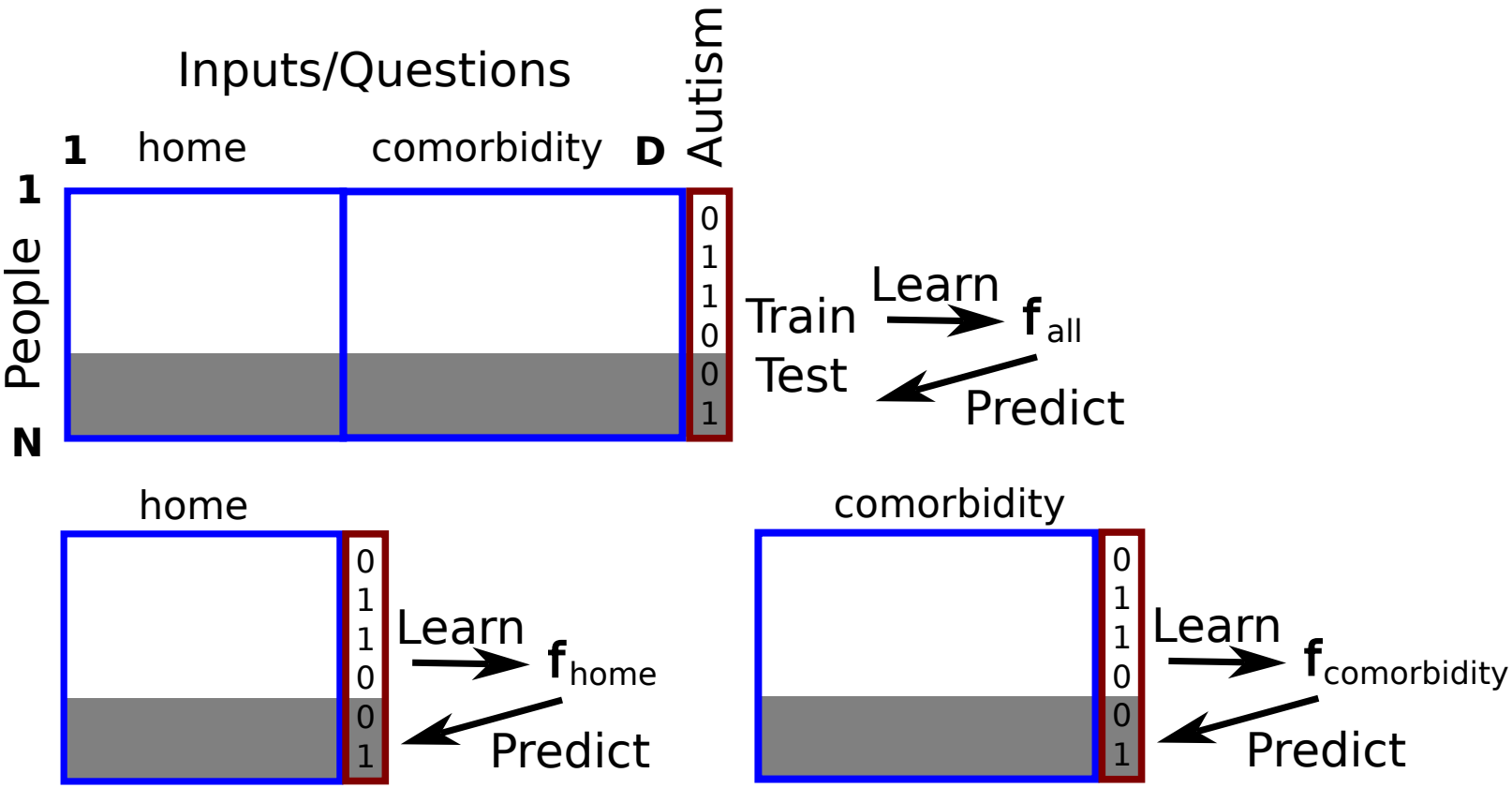
Similarity/difference between years

Discussion and Conclusions

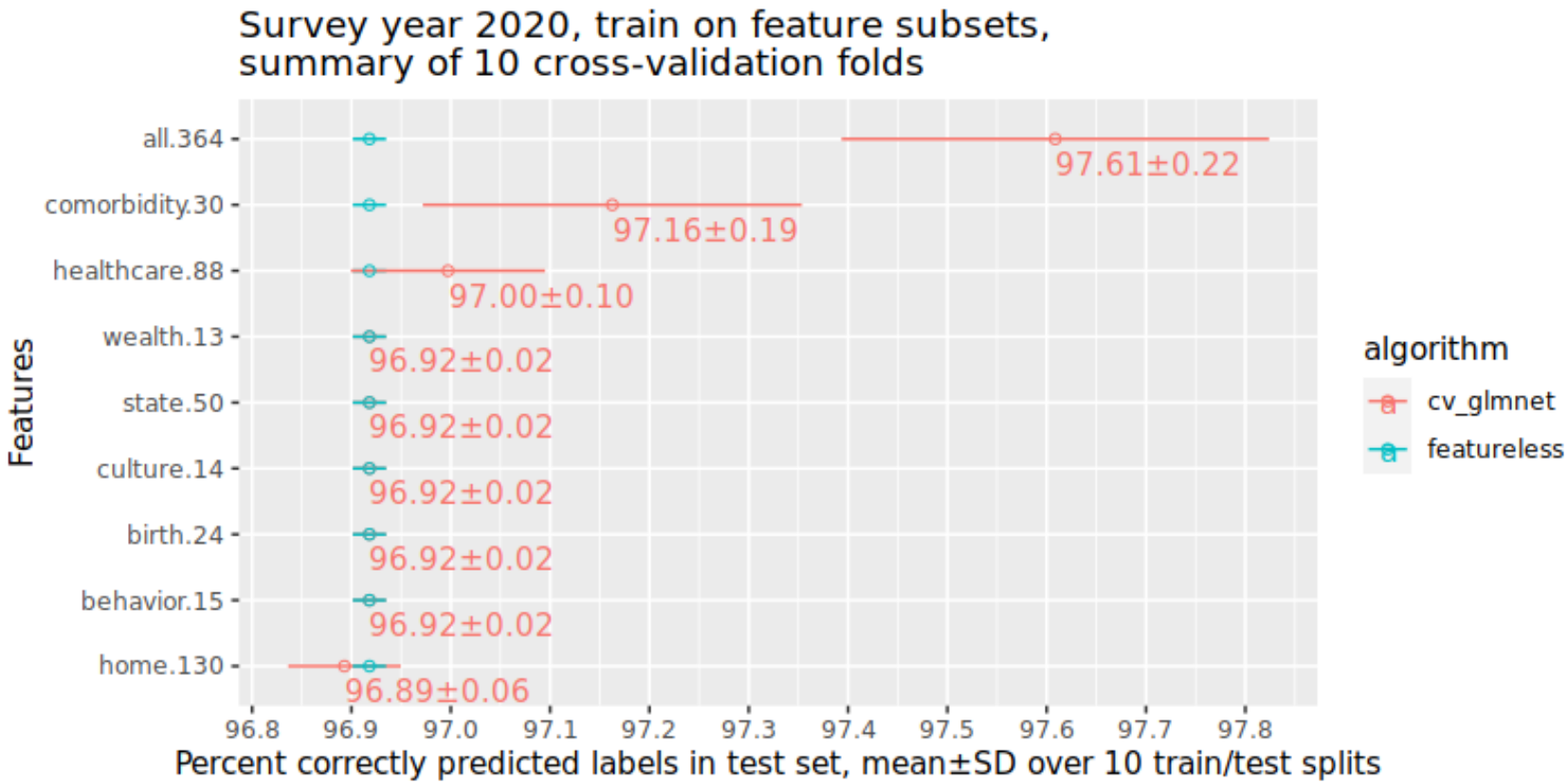
Column categorization

TODO

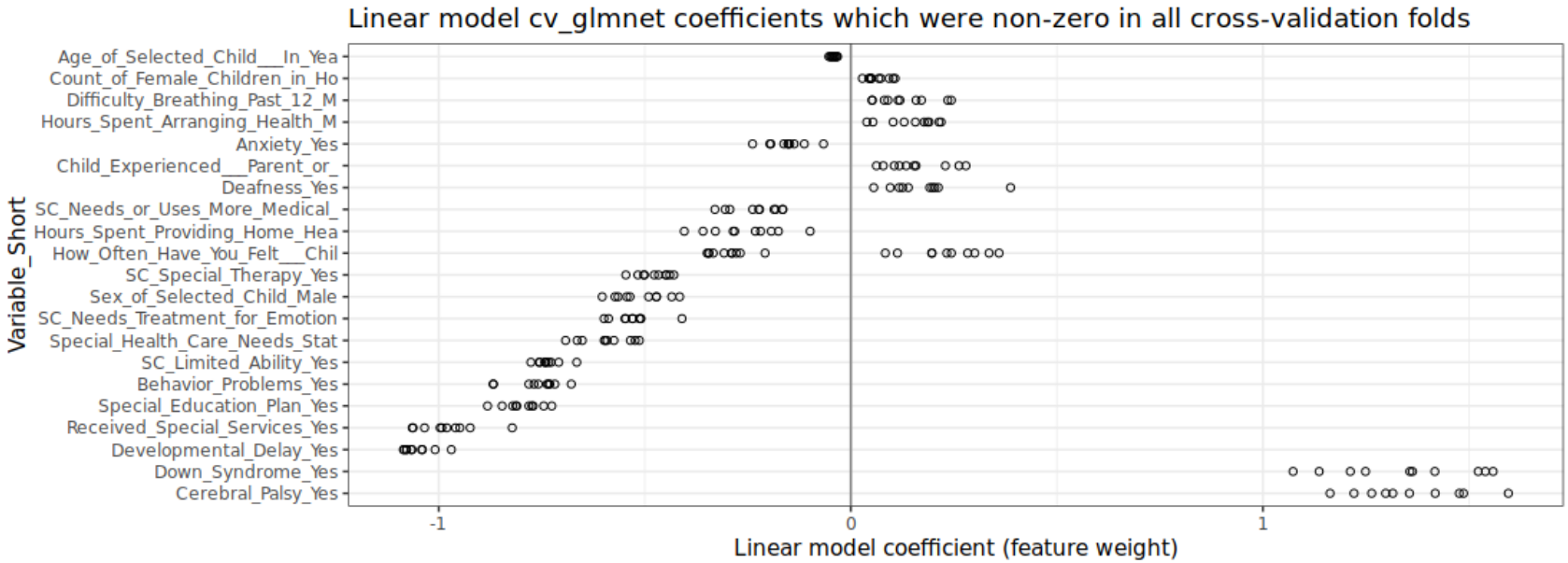
Cross-validation for feature importance



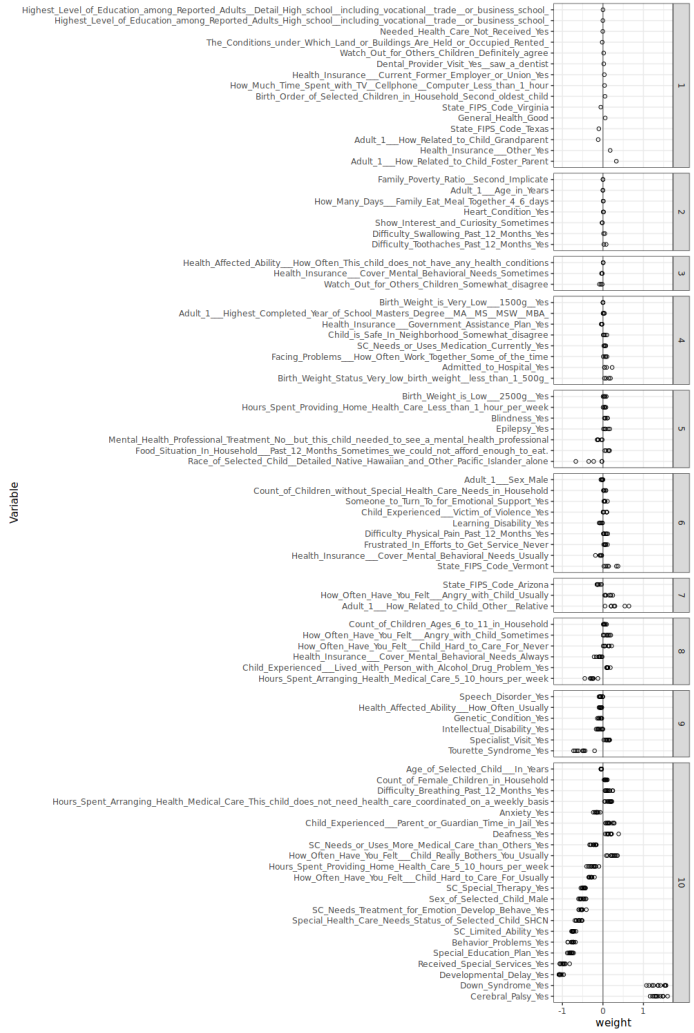
Cross-validation for feature importance



Cross-validation for feature importance



Cross-validation for feature importance



View full figure online, `https://github.com/tdhock/2024-01-ml-for-autism/blob/main/download-nsch-mlr3batchmark-registry-glmnet-coef.png`

Data pre-processing

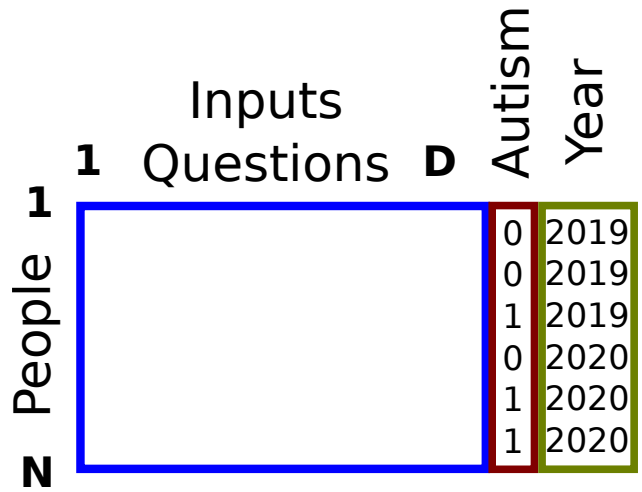
Prediction accuracy in a given year

Model interpretation / feature selection

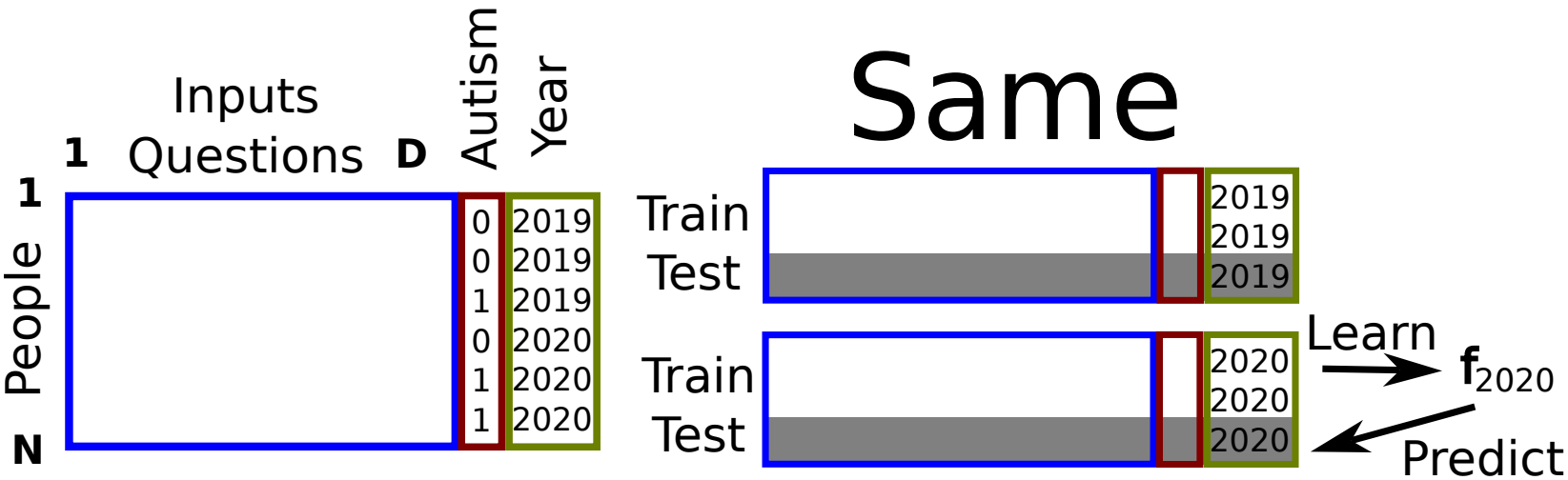
Similarity/difference between years

Discussion and Conclusions

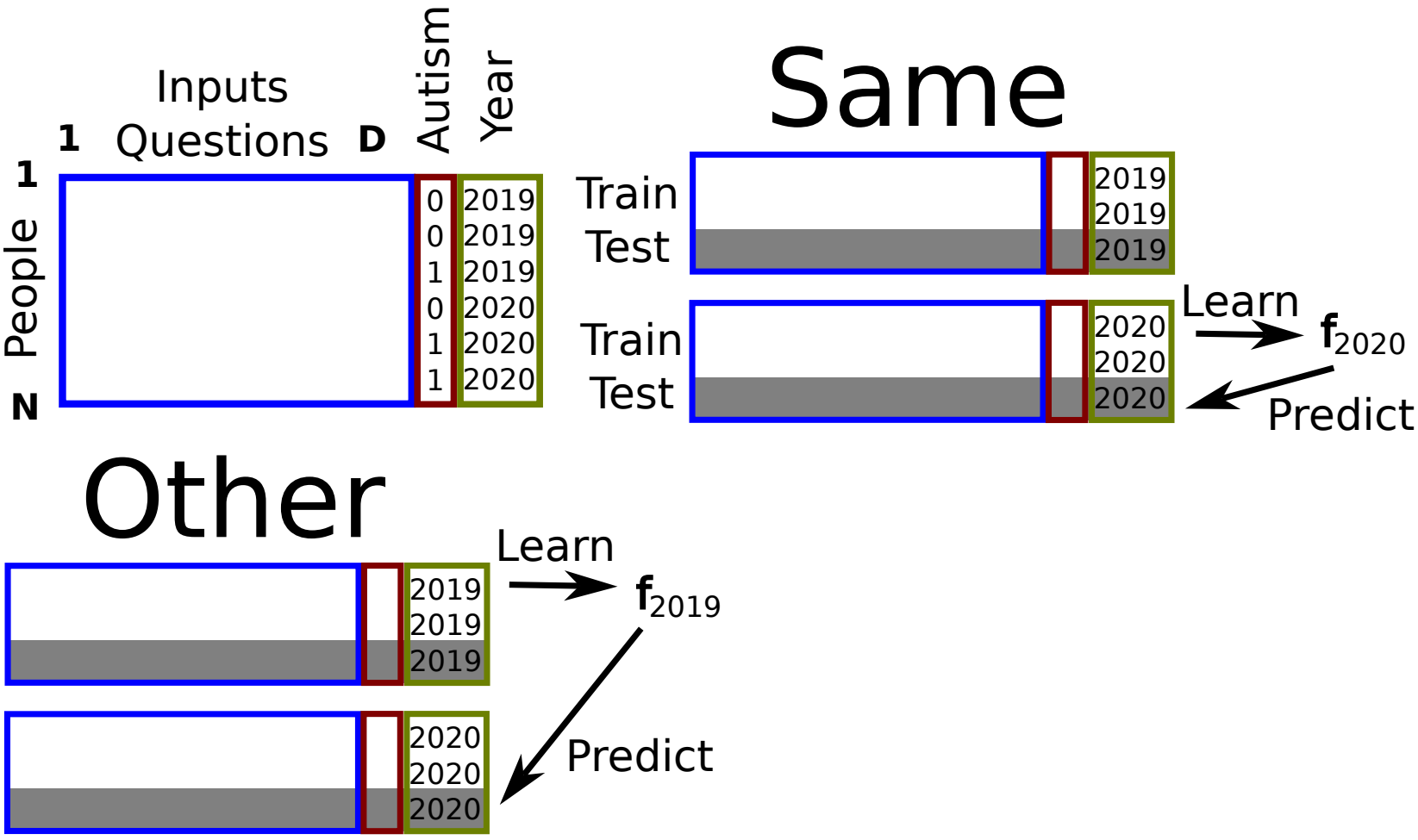
Cross-validation for determining similarity between years



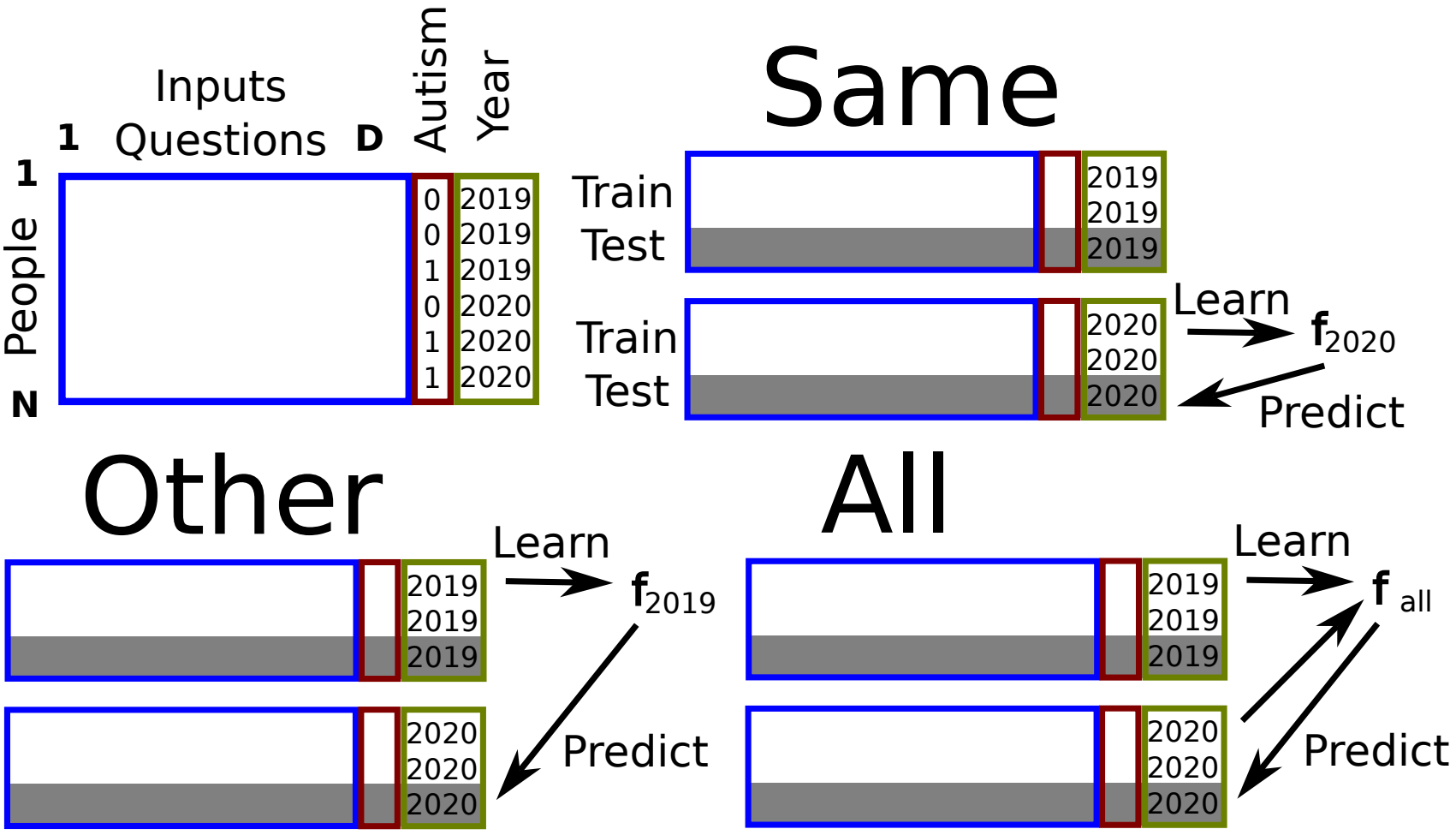
Cross-validation for determining similarity between years



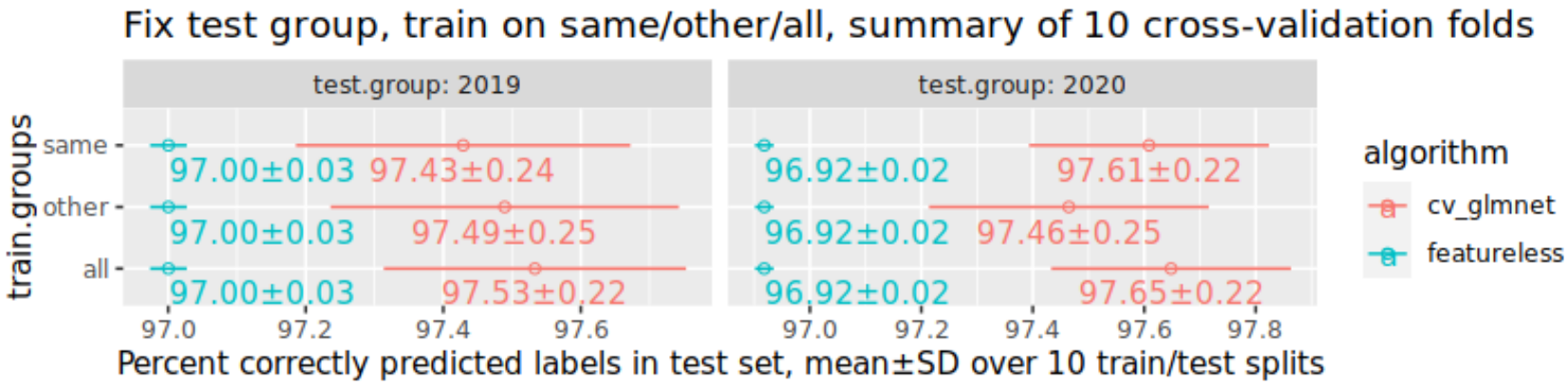
Cross-validation for determining similarity between years



Cross-validation for determining similarity between years



Cross-validation for determining similarity between years



- 18,202 rows in 2019, whereas 27,808 in 2020.

Data pre-processing

Prediction accuracy in a given year

Model interpretation / feature selection

Similarity/difference between years

Discussion and Conclusions

Discussion and Conclusions

- Often we want to know if we have similar or different patterns in different groups (train on one year, predict on another).
- Cross-validation can be used to determine the extent to which we can train on one group, and accurately predict on another.
- Machine learning algorithms like L1 regularized linear models (LASSO/cv_glmnet) are additionally interpretable in terms of which features are used for prediction (can be compared between models trained on different groups).
- Free/open-source software available: mlr3resampling R package on CRAN and `https://github.com/tdhock/mlr3resampling`
- Let's collaborate! Contact: toby.hocking@nau.edu, toby.hocking@r-project.org