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 Autism variable (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

- ▶ Download NSCH data from public Census web site.
- ▶ For each year, keep all columns with less than 10% missing values, then remove all rows with at least one missing value.
- ▶ One-hot recoding of categorical variables (create 0/1 dummy/indicator variable for each value).
- ▶ Then keep only columns in common between both years: result is 46,010 rows and 366 columns. Details:

year	data.type	nrow	ncol	questions	%Autism	%rowsNA	%colsNA
2019	raw	29433	443	443	2.9615	100.0000	90.0677
2019	processed	18202	377	187	2.9997	0.0000	0.0000
2020	raw	42777	443	443	2.9758	100.0000	90.0677
2020	processed	27808	373	185	3.0818	0.0000	0.0000

Data source:
`http://www2.census.gov/programs-surveys/nsch/datasets/2019/nsch_2019_topical_Stata.zip` nsch_2019_topical.dta, nsch_2019_topical.do
`http://www2.census.gov/programs-surveys/nsch/datasets/2020/nsch_2020_topical_Stata.zip` nsch_2020_topical.dta, nsch_2020_topical.do

One-hot encoding of categorical variables

Sometimes called dummy/indicator variables in statistics.
For each value, we create a new column with 0/1 values.
For example, from nsch_2020_topical.do

```
label var k4q24_r  "Specialist Visit"
label define k4q24_r_lab  1  "Yes"
label define k4q24_r_lab  2  "No, but this child needed
 to see a specialist", add
label define k4q24_r_lab  3  "No, this child did not
 need to see a specialist", add
```

code above means there is a column named k4q24_r in
nsch_2020_topical.dta, with values 1, 2, 3.
In our analysis we use a one-hot encoding, which means deleting
that column, and creating two 0/1 columns:
Specialist Visit=Yes and
Specialist Visit=No, but this child needed to see a specialist
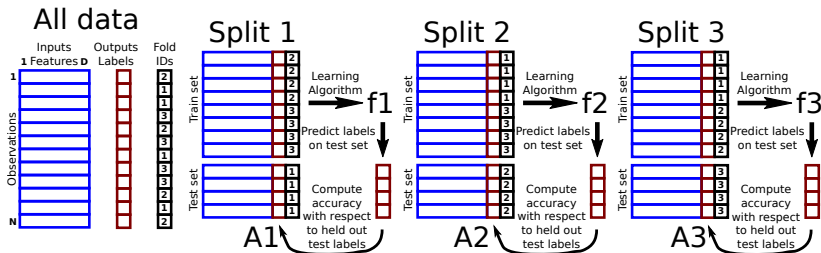
Data pre-processing

Prediction accuracy in a given year

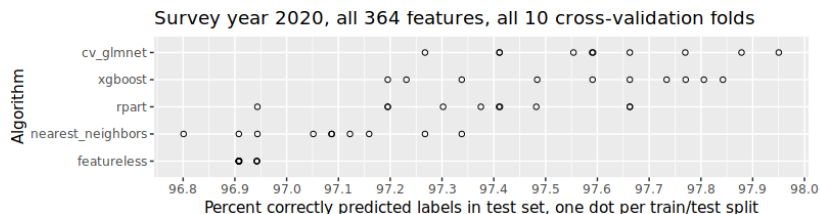Model interpretation / feature selection

Similarity/difference between years
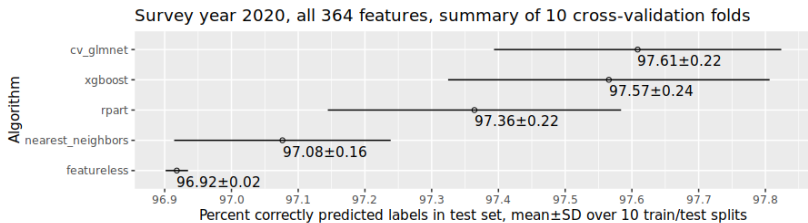
Discussion and Conclusions

K-fold cross-validation: a standard algorithm used to estimate the prediction accuracy in machine learning

10-fold cross-validation for comparing learning algorithms

Survey year 2020, all 364 features, all 10 cross-validation folds

Summarize 10 folds with mean and standard deviation

Survey year 2020, all 364 features, summary of 10 cross-validation folds

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 $f(x)$.
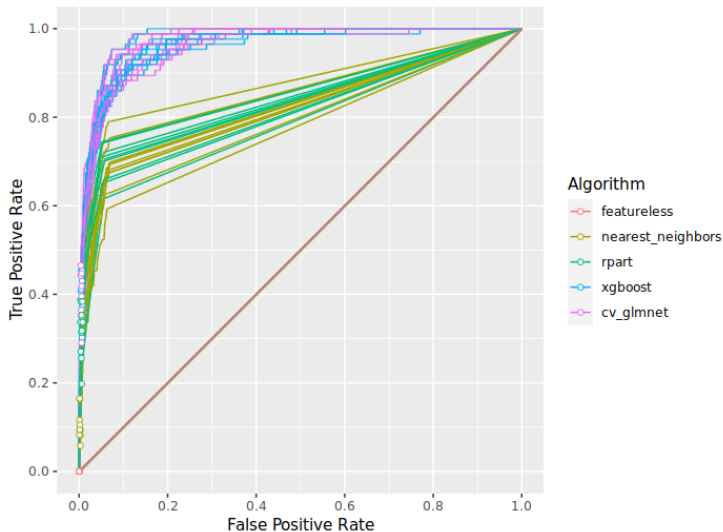
ROC curves show all tradeoffs between TPR and FPR

Survey year 2020, all 364 features,
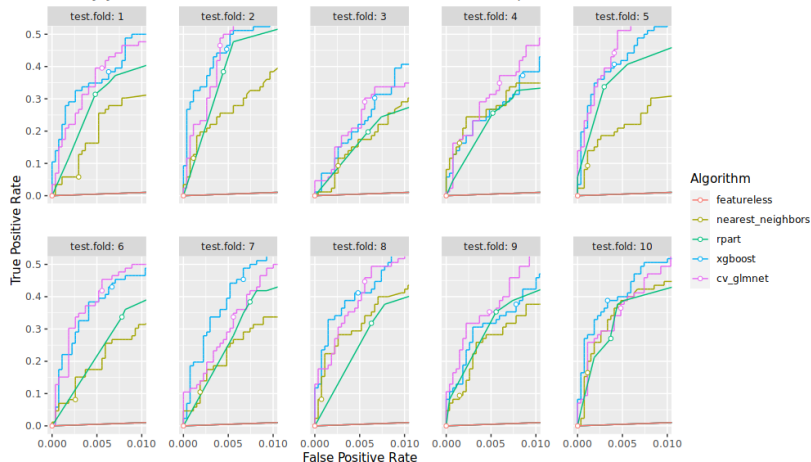One ROC curve per cross-validation fold
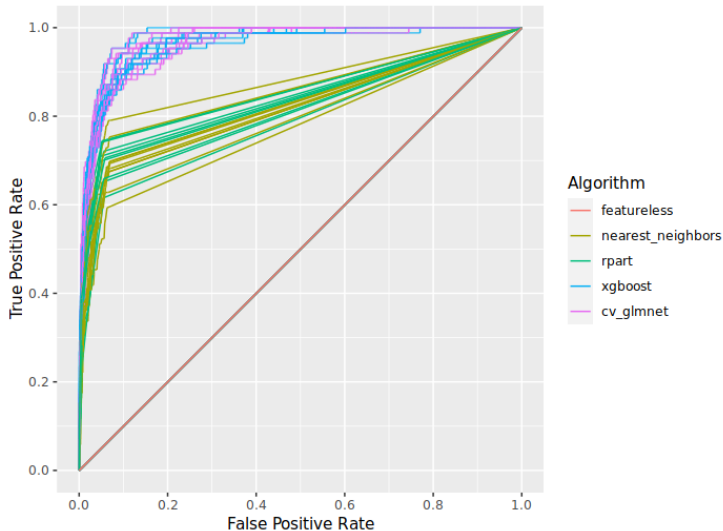
Default prediction threshold can be viewed as a dot

Survey year 2020, all 364 features,
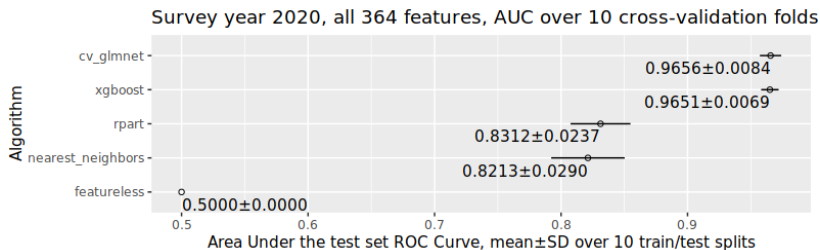One ROC curve per cross-validation fold

Default prediction threshold can be viewed as a dot

Survey year 2020, all 364 features, zoom to show FPR/TPR of predictions

Algorithm
- featureless
- nearest_neighbors
- rpart
- xgboost
- cv_glmnet

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

Survey year 2020, all 364 features, AUC over 10 cross-validation folds

Algorithm	
cv_glmnet	0.9656±0.0084
xgboost	0.9651±0.0069
rpart	0.8312±0.0237
nearest_neighbors	0.8213±0.0290
featureless	0.5000±0.0000

Area Under the test set ROC Curve, mean±SD over 10 train/test splits

Data pre-processing

Prediction accuracy in a given year

Model interpretation / feature selection

Similarity/difference between years

Discussion and Conclusions

Column categorization

Each input/feature column was assigned a category.

```
column_name,category
survey_year,
Autism,
State FIPS Code=Alabama,state
...
Number of Children in Household=1,home
...
Sex of Selected Child=Male,birth
...
Deafness=Yes,comorbidity
...
```
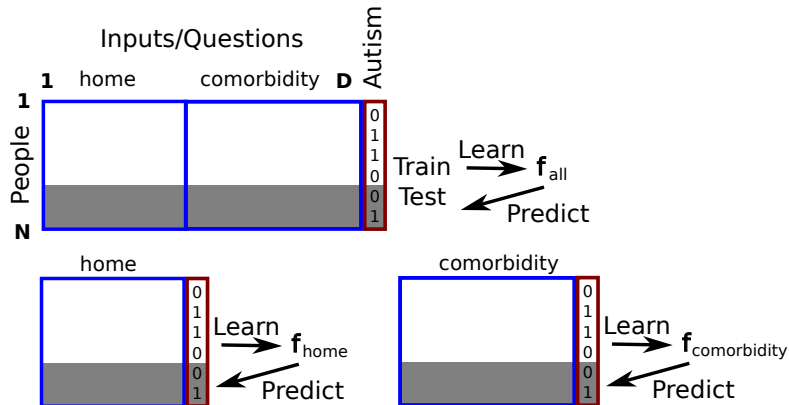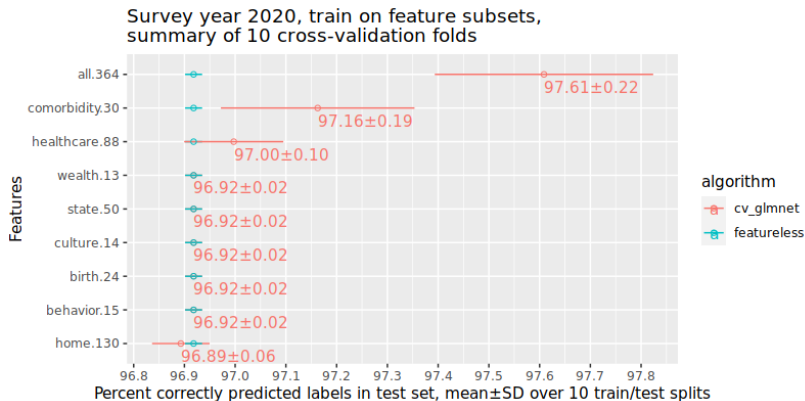
		behavior	birth	comorbidity
		2	15	24	30

	culture	healthcare	home	state
	14	88	130	50

	wealth
	13

Cross-validation for category importance

Cross-validation for category importance

Survey year 2020, train on feature subsets, summary of 10 cross-validation folds. Percent correctly predicted labels in test set, mean±SD over 10 train/test splits.
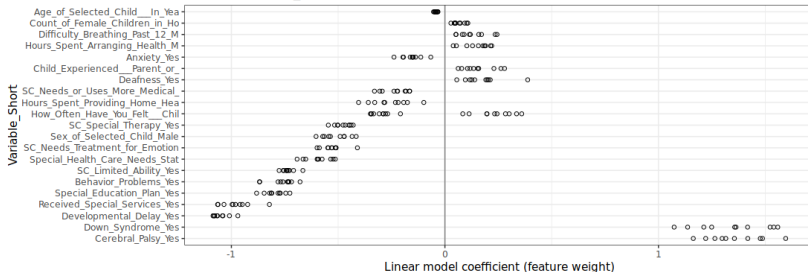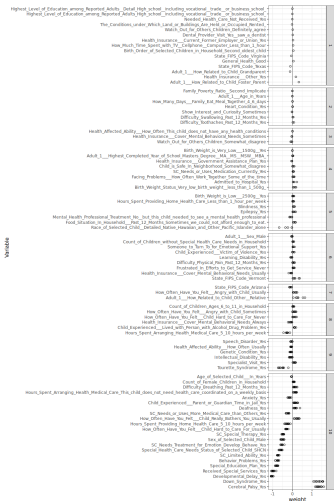
TODO equation

- Positive coefficients mean larger values of that feature are more likely to TODO

Linear model coefficient / feature importance

Linear model cv_glmnet coefficients which were non-zero in all cross-validation folds

Linear model coefficient / 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
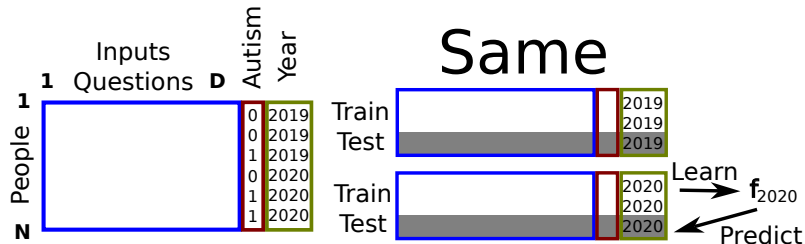
Model interpretation / feature selection

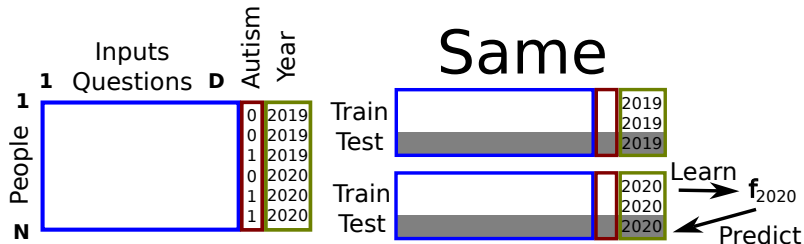Similarity/difference between years

Discussion and Conclusions

Cross-validation for determining similarity between years

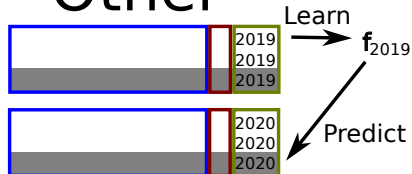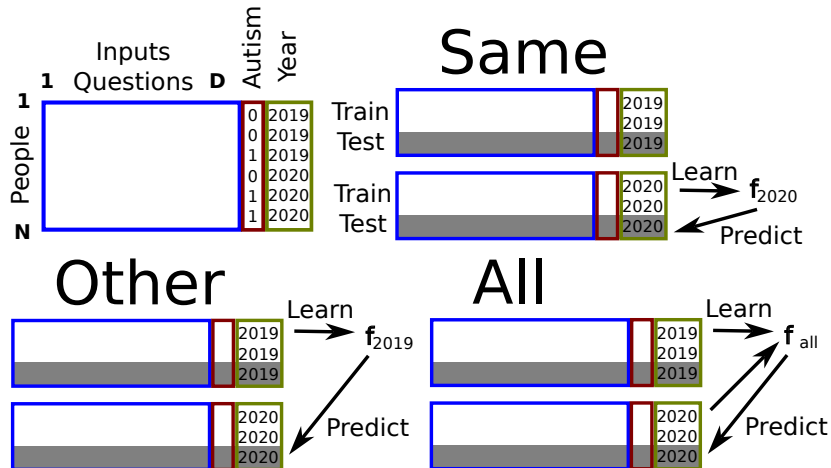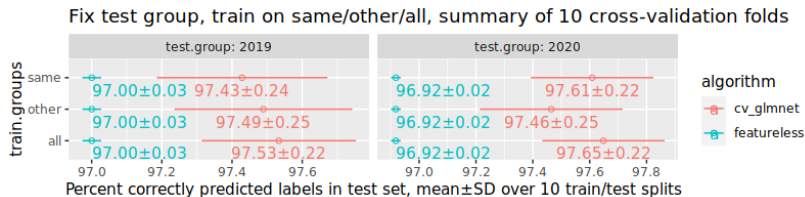# Cross-validation for determining similarity between years

Cross-validation for determining similarity between years

Cross-validation for determining similarity between years

Cross-validation for determining similarity between years

Fix test group, train on same/other/all, summary of 10 cross-validation folds

Percent correctly predicted labels in test set, mean±SD over 10 train/test splits

▶ 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