Predictive Modeling

PART-5

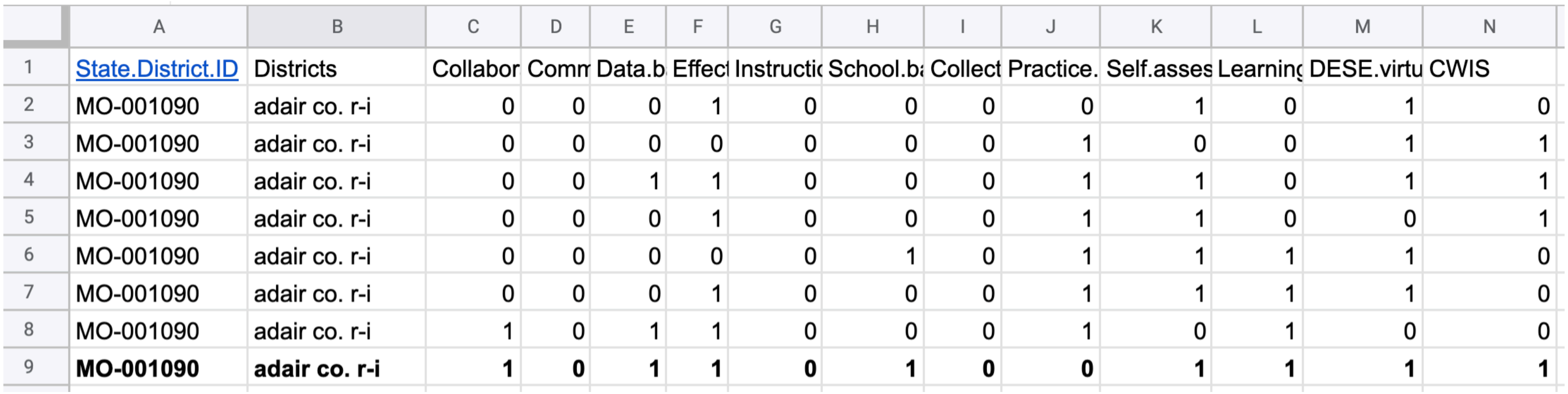
Districts Aggregate

Coaching Logs

Rows = 192 & Columns = 14

Rows = 5924 & Columns = 58

- Each line of the coaching logs is an interaction between a school and a Coach.
- Doesn't have building level or school level data unlike NCES, CWIS. (Need for aggregates)
- Identifier for merging tables - **State.District.ID**
- **MO-001090(adair co. r-i) -> (multiple schools, building in CWIS,NCES) -> we need an aggregate**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	State.District.ID	Districts	Collabor	Comm	Data.b	Effect	Instructi	School.b	Collect	Practice.	Self.asses	Learning	DESE.virtu	CWIS
2	MO-001090	adair co. r-i	0	0	0	1	0	0	0	0	1	0	1	0
3	MO-001090	adair co. r-i	0	0	0	0	0	0	0	1	0	0	1	1
4	MO-001090	adair co. r-i	0	0	1	1	0	0	0	1	1	0	1	1
5	MO-001090	adair co. r-i	0	0	0	1	0	0	0	1	1	0	0	1
6	MO-001090	adair co. r-i	0	0	0	0	0	1	0	1	1	1	1	0
7	MO-001090	adair co. r-i	0	0	0	1	0	0	0	1	1	1	1	0
8	MO-001090	adair co. r-i	1	0	1	1	0	0	0	1	0	1	0	0
9	**MO-001090**	**adair co. r-i**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**1**	**1**	**1**

CWIS

Rows = 80267 & Columns = 106

- Survey responses
- Also, if a district met with a coaching team, but didn't fill out the CWIS survey, you'll find the district in the Coaching Logs but not in the CWIS survey data.
- Building/School level data.
- 867 unique school IDs

NCES

Rows = 2456 & Columns = 26

- National Common Core data that includes descriptions of the buildings (**e.g., Free/Reduced lunch rate, student:teacher ratio, rural/urban, etc.**)
- Building/School level data.
- 2456 nces unique school IDs

CWIS + NCES

Rows = 78558 & Columns = 123

- Merging CWIS,NCES based on State.School.ID

- **1709 missing rows are because :**
- There are only 839 common school IDs between CWIS and NCES.
- NCES data doesn’t have all the school IDs

CWIS + NCES

Rows = 78558 & Columns = 123

- Building/School level data.
- To integrate with Districts we use District IDs here
- 229 unique District IDs

Districts Aggregate

Rows = 192 & Columns = 14

- District level Aggregate data for schools.
- To integrate with CWIS&NCES we use District IDs here
- 192 unique District IDs

CWIS + NCES +Districts Aggregate

Rows = 77375 & Columns = 136

- **Merging CWIS,NCES based on State.District.ID**

- **1183 missing rows are because :**
- There are only 183 common Districts IDs between CWIS+NCES and District Aggregates.
- Indicating few of the districts weren't involved in Coaching.

Problems

- Need building/School level data in Coaching Logs to get Improving Instruction school wise.
- Consistency metric per month?
- How to combine CWIS + Coaching for ETL average in this case?

Baseline Models

- Simple Models that provide reasonable results and requires less expertise or time to build.
- Baselines predictions are independent of the inputs.
- Why Baseline?

Hypothesis: If there is any relation between inputs and outputs

- If the baselines do better than our models(regression or classification) it's mean we can assume the inputs has no relation with the outputs.

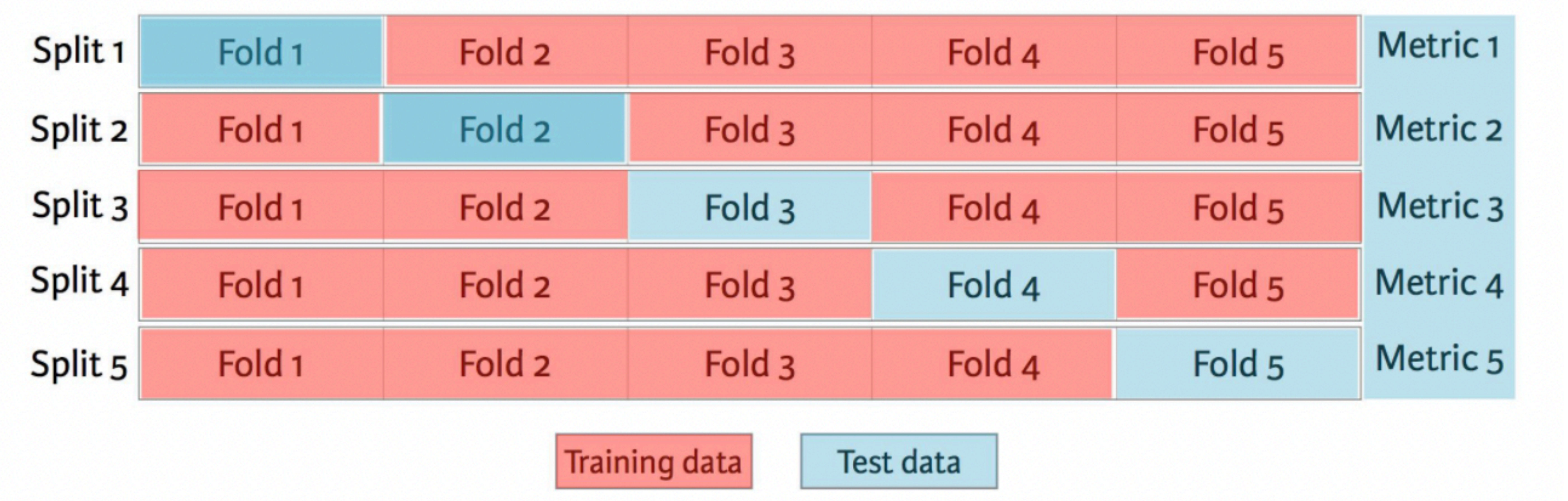
<u>Baselines Used:</u>

- **L0 - Classification Baseline -**> **frequency**
- **L1 - Regression Baseline -**> **median**
- **L2 - Regression Baseline -**> **mean**

Linear Regression

cvglmnet - cross validation + (Lasso-L1)

- Regularization works by penalizing the magnitude of coefficients of the features.
- L1 tends to shrink coefficients to zero.
- L1 is therefore useful for feature selection, as we can drop any variables associated with coefficients that go to zero



Split 1	🟦 Fold 1	🟥 Fold 2	🟥 Fold 3	🟥 Fold 4	🟥 Fold 5	Metric 1
Split 2	🟥 Fold 1	🟦 Fold 2	🟥 Fold 3	🟥 Fold 4	🟥 Fold 5	Metric 2
Split 3	🟥 Fold 1	🟥 Fold 2	🟦 Fold 3	🟥 Fold 4	🟥 Fold 5	Metric 3
Split 4	🟥 Fold 1	🟥 Fold 2	🟥 Fold 3	🟦 Fold 4	🟥 Fold 5	Metric 4
Split 5	🟥 Fold 1	🟥 Fold 2	🟥 Fold 3	🟥 Fold 4	🟦 Fold 5	Metric 5

🟥 Training data 🟦 Test data

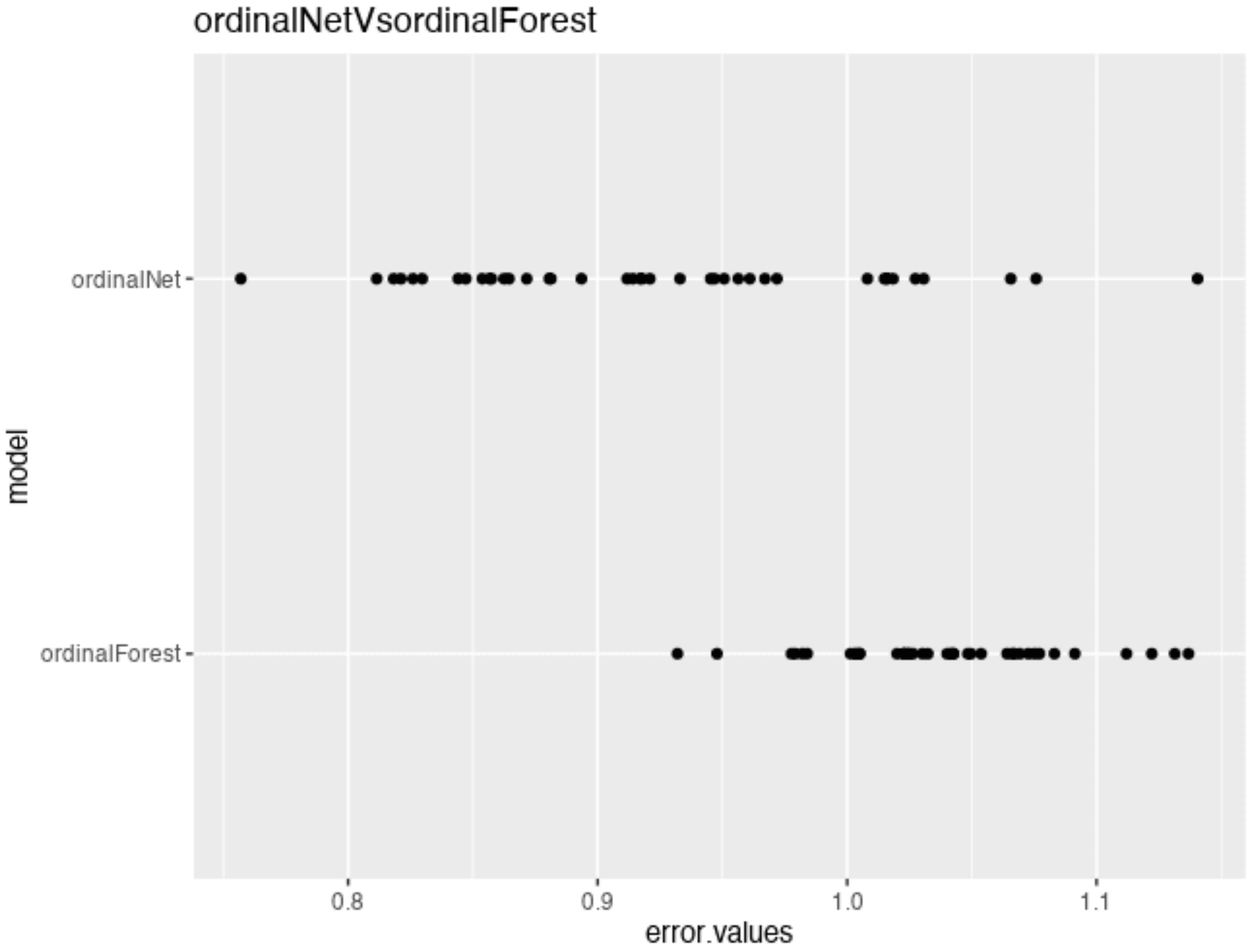
5-fold cross validation (image credit)

OrdinalNet

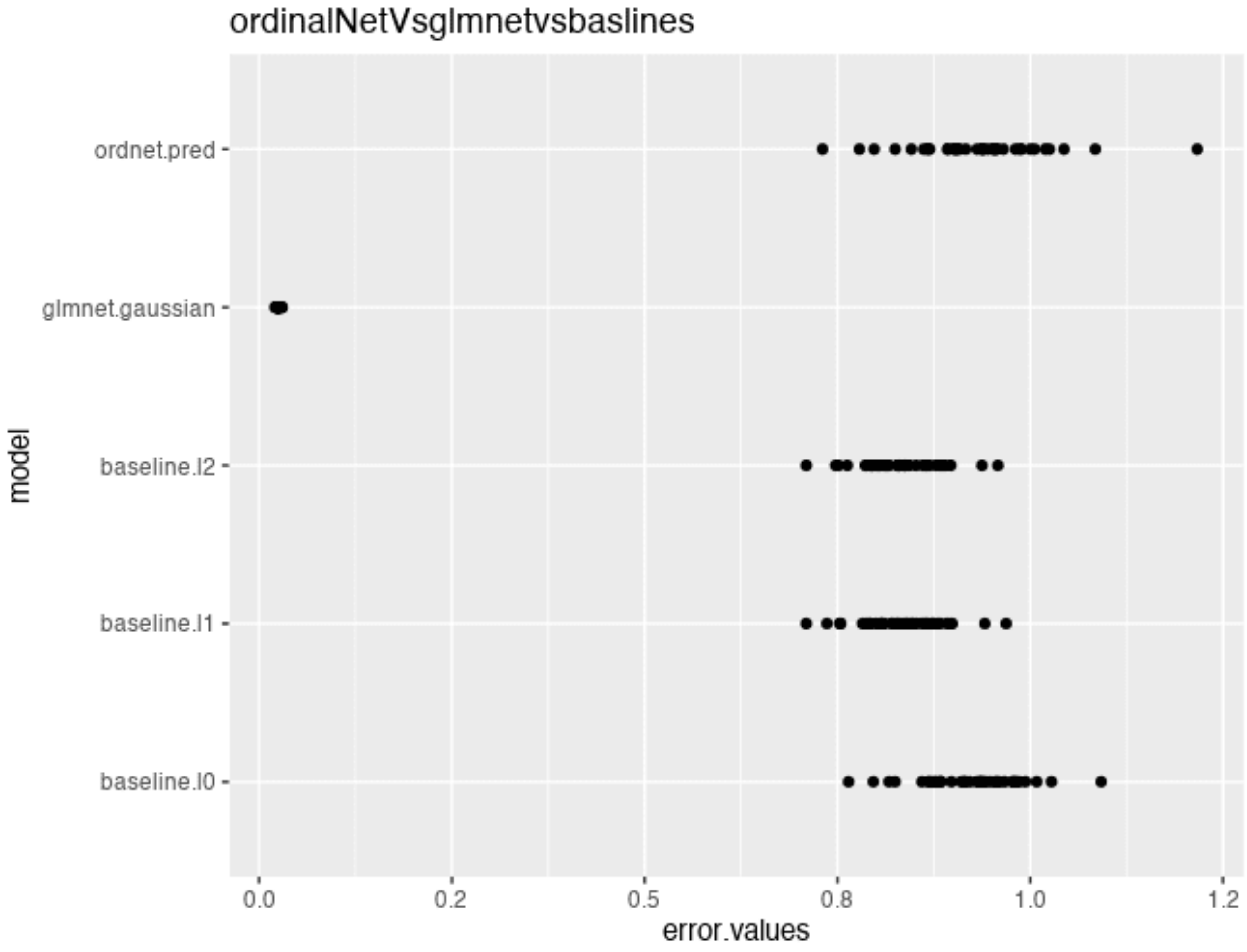
- In statistics, ordinal regression (also called "ordinal classification") is a type of regression analysis used for predicting an ordinal variable, with 'ordered' multiple categories and independent variables.
- Ordinal Net fits ordinal regression models with elastic net penalty.

Ordinal Forest

- The ordinal forest (OF) method allows ordinal regression with high-dimensional and low-dimensional data.
- Moreover, by means of the (permutation-based) variable importance measure of OF, it is also possible to rank the covariates with respect to their importance in the prediction of the values of the ordinal target variable.



OrdinalNet Vs glmnet Vs Baselines


ordinalNetVsglmnetvsbaslines
model
ordnet.pred
glmnet.gaussian
baseline.l2
baseline.l1
baseline.l0
error.values