Predictive Modeling

Akhila Chowdary Kolla, Shadmaan Hye
Joint work with Dr.Toby Dylan Hocking

September 29, 2021

Outline

Our work so far

- Implementation glmnet
- Implementation baselines
- Comparison of glmnet with baseline

DataSet Provided

- Coaching logs (Rows = 5924 Columns = 58)
- CWIS data – includes the outcome variable (ETL Average) (Rows = 80267 Columns = 106)
- NCES District information (Rows = 2456 Columns = 26)

Combination of the 3 datasets

- Combined dataset (Rows = 5610993, Columns = 26)
- Missing values

Modified Combination of the 3 datasets

- Removed the rows containing more missing values
- Analysed which columns are required for answering the questions
  Q2. To what extent do demographic or community factors correlate with the rate of improving instruction?
  Q3. What attributes of external support (externally provide training, coaching, DESE support) influence the rate of improving instruction? What are the conditions that cause the contribution of these variables to vary?
- Substituted "Yes" to 1, "No" to 0 for survey data in the files
- New created dataset with no mising values (Rows = 25547 Columns = 38)

Baselines in Machine Learning

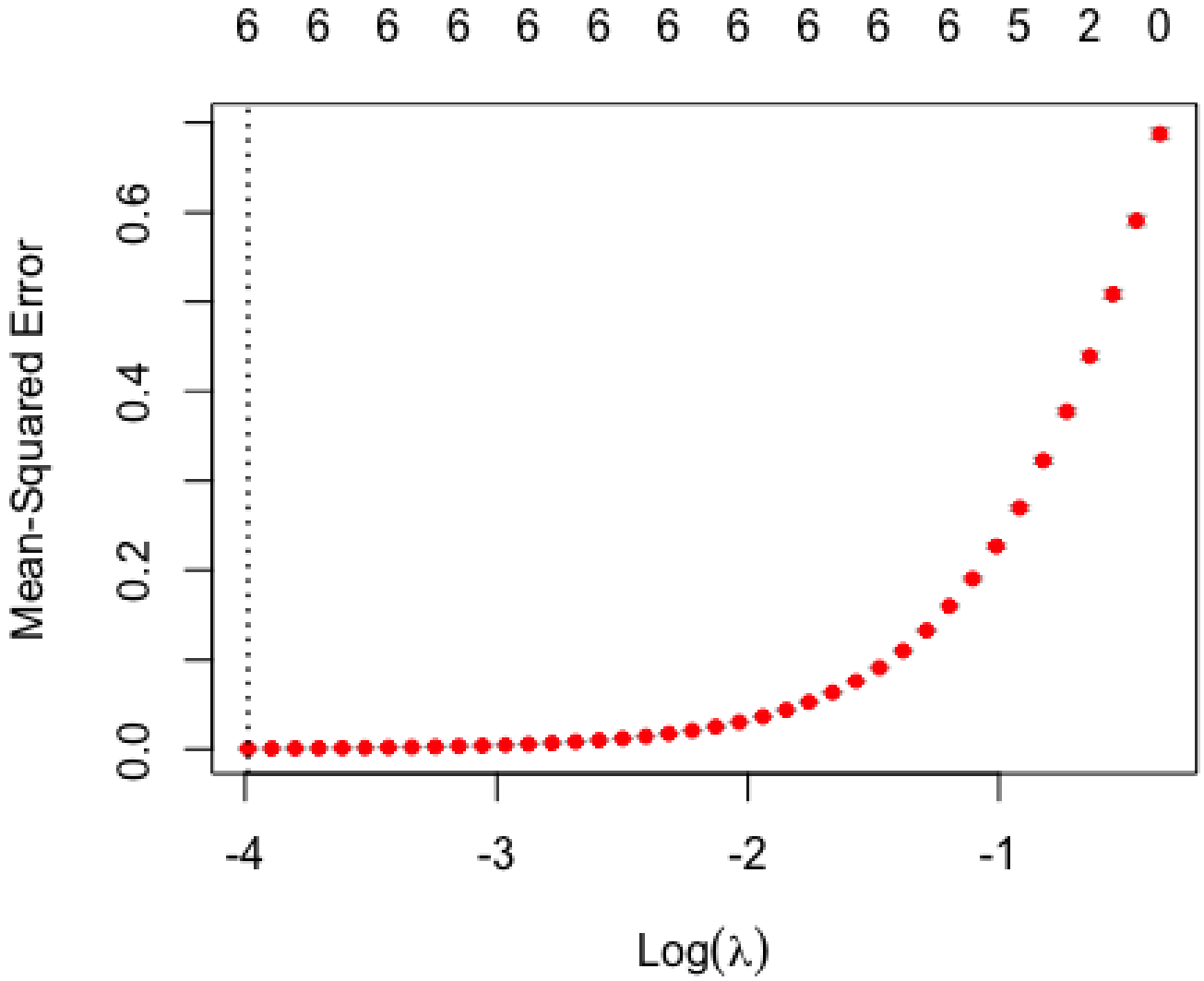
- A baseline is a method that uses heuristics, simple summary statistics, randomness, or machine learning to create predictions for a dataset.
- We can use these predictions to measure the baseline's performance (e.g., accuracy).
- We can compare the model performance with the baseline's performance

cvglmnet: Cross-validation for glmnet

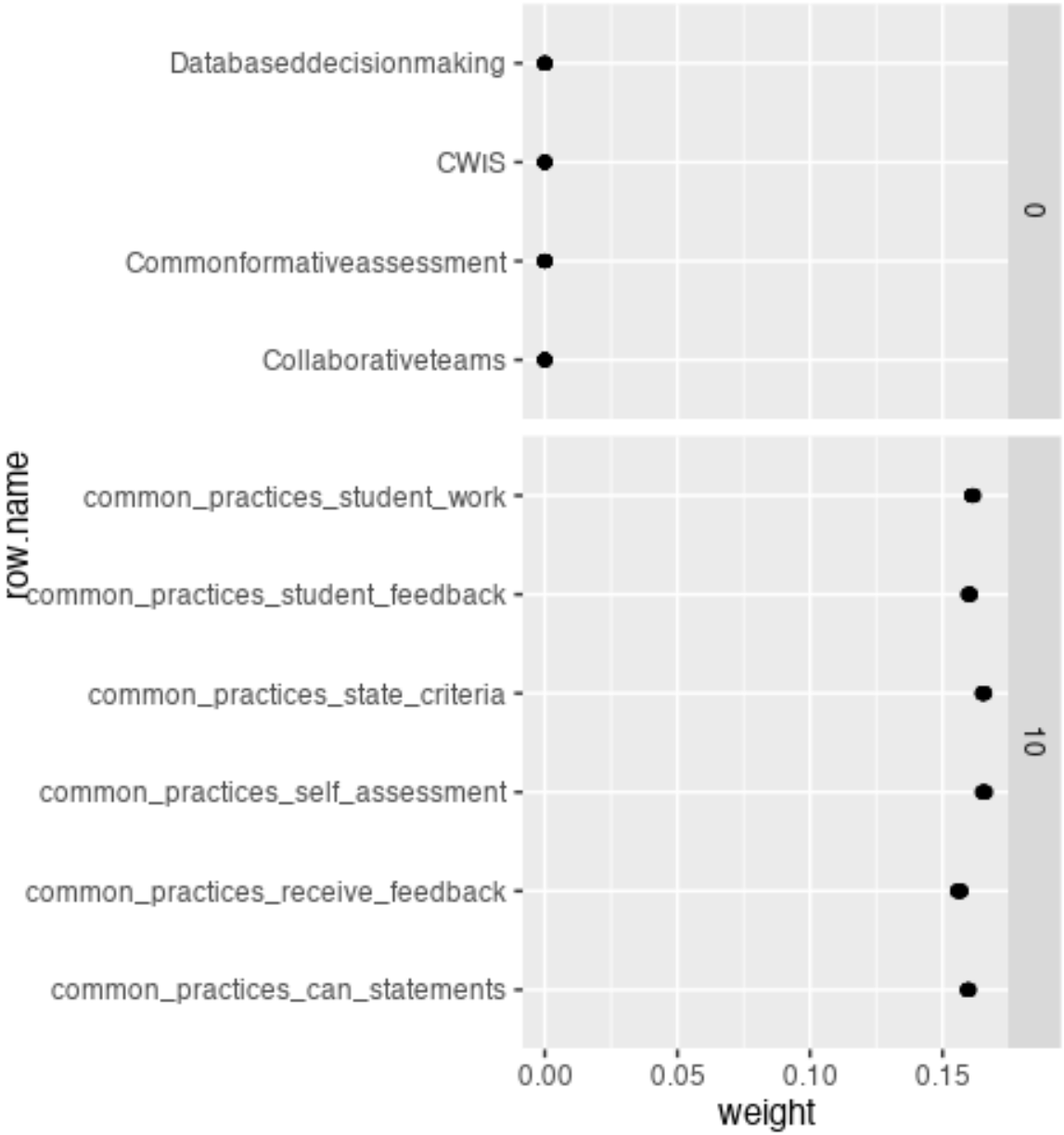
cvglmnet?

- cvglmnet gives optimal Lambda value instead of a set of Lambda values
- easier to use the Lambda function

Results from cvglmnet on the dataset



For multiple 10.folds



Evaluation Metrics

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

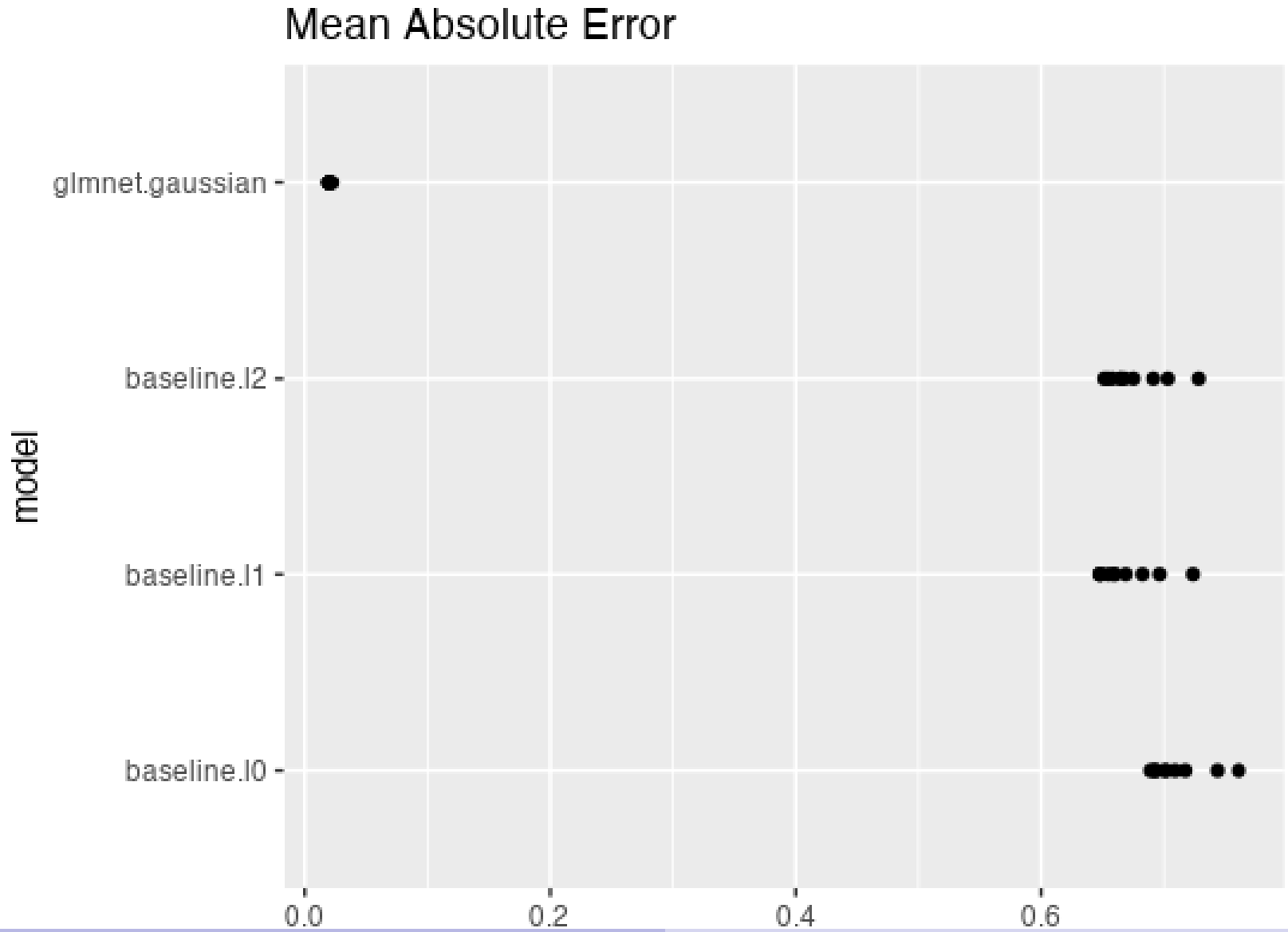
Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

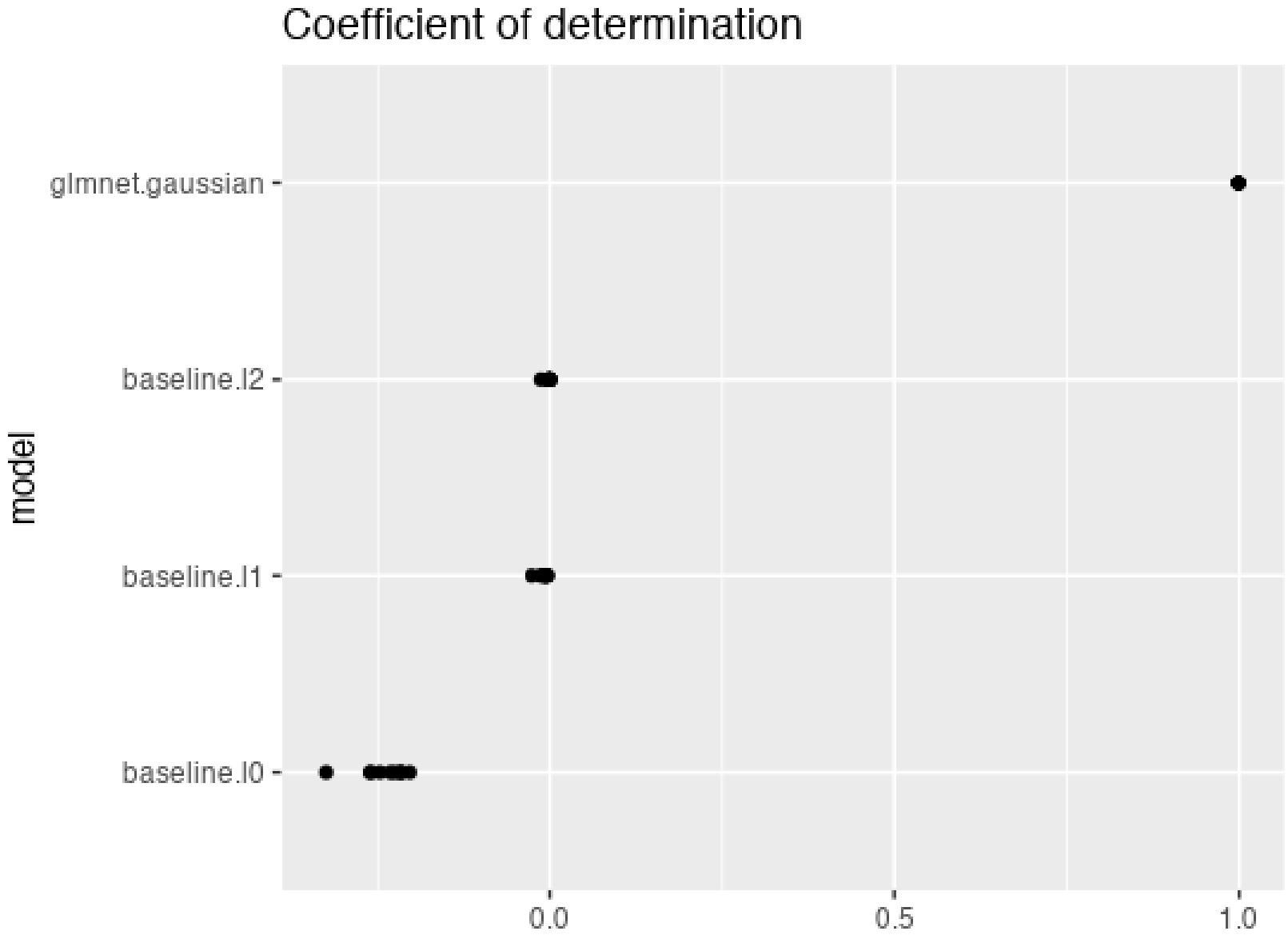
For multiple 10.folds - MAE

MAE represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.



For multiple 10.folds - R^2

$R\hat{2}$ represents the coefficient of how well the values fit compared to the original values.



Our future goals

- Comparisons between Ordinal net and Ordinal Forest to understand which model works better

Plots

- Training the dataset with a) keeping missing rows, b) filling missing values
- Nonlinear Function (Random Forest) better than Linear Function (glmnet)? OrdinalNet vs Ordinal Forest