Cross Validation

Improving How We Evaluate our Machine Learning Models

How have we been evaluating our models?

- What’s the process?

How have we been evaluating our models?

- Train-Test Split
  - Make a training set to teach the models the pattern in the data
  - Make a testing set to evaluate the model

Why have separate Train and Test sets?

- What do you think?

Why have separate Train and Test sets?

- To prevent **overfitting**
- Having the Testing set be included in the Training set is like giving students the answer key to a test
  - Of course they do well!
  - You don’t correctly evaluate their understanding

Pros of Train Test Split

- What do you think?

Pros of Train Test Split

- You can evaluate your models fairly.
- It’s easy to do.
- It’s easy to change how much data is set aside for the testing set

Cons of Train Test Split

- What do you think?

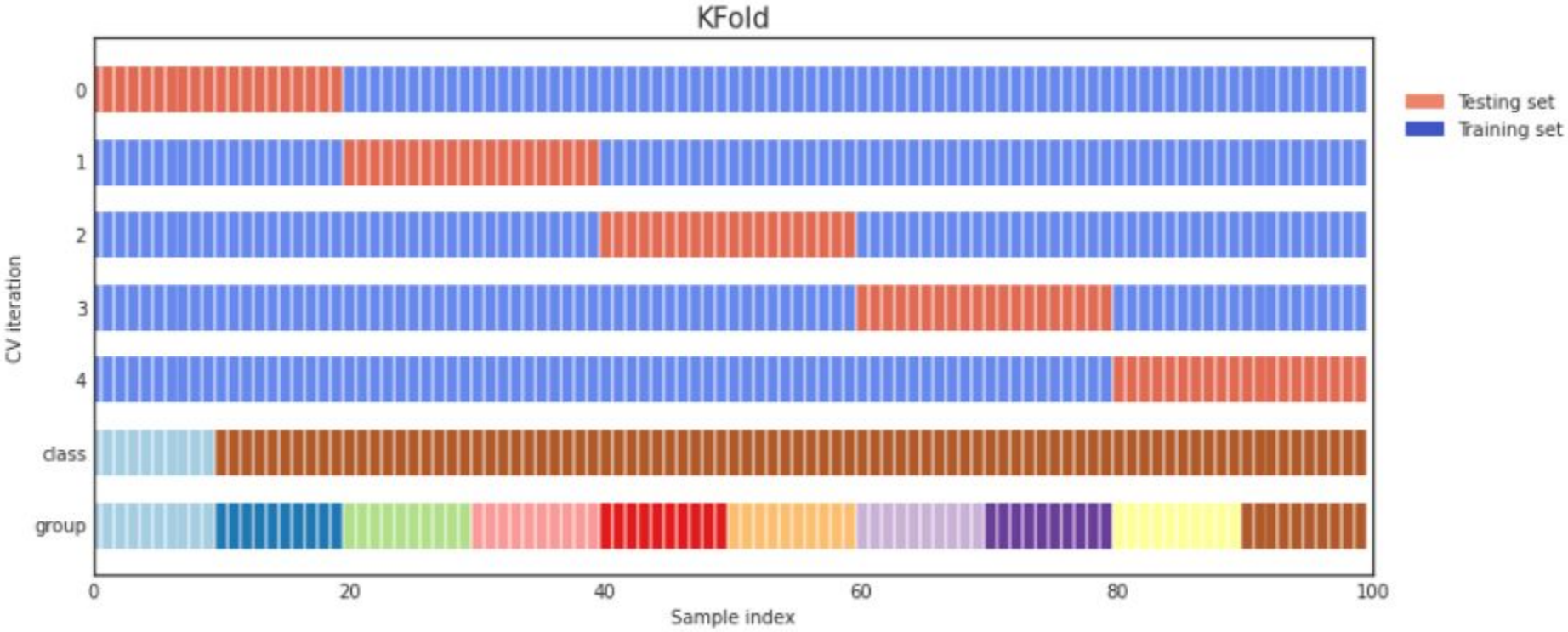
Cons of Train Test Split

- Sometimes, your models good or bad by chance- due to the random split of train/testing set
- If your data is unbalanced (ex: cancer diagnoses), often there are very few positive cases in a testing set, so you do not get the best view of the performance of a model

Insert Cross Validation!

- You split the dataset into folds. 1 fold is set aside for testing/validation, the rest are set aside for training.
- Once you train and evaluate your model, you do it again from scratch, but with a different fold as the testing set
- Often called KFold where K is the number of folds

How it works (ignore group row)



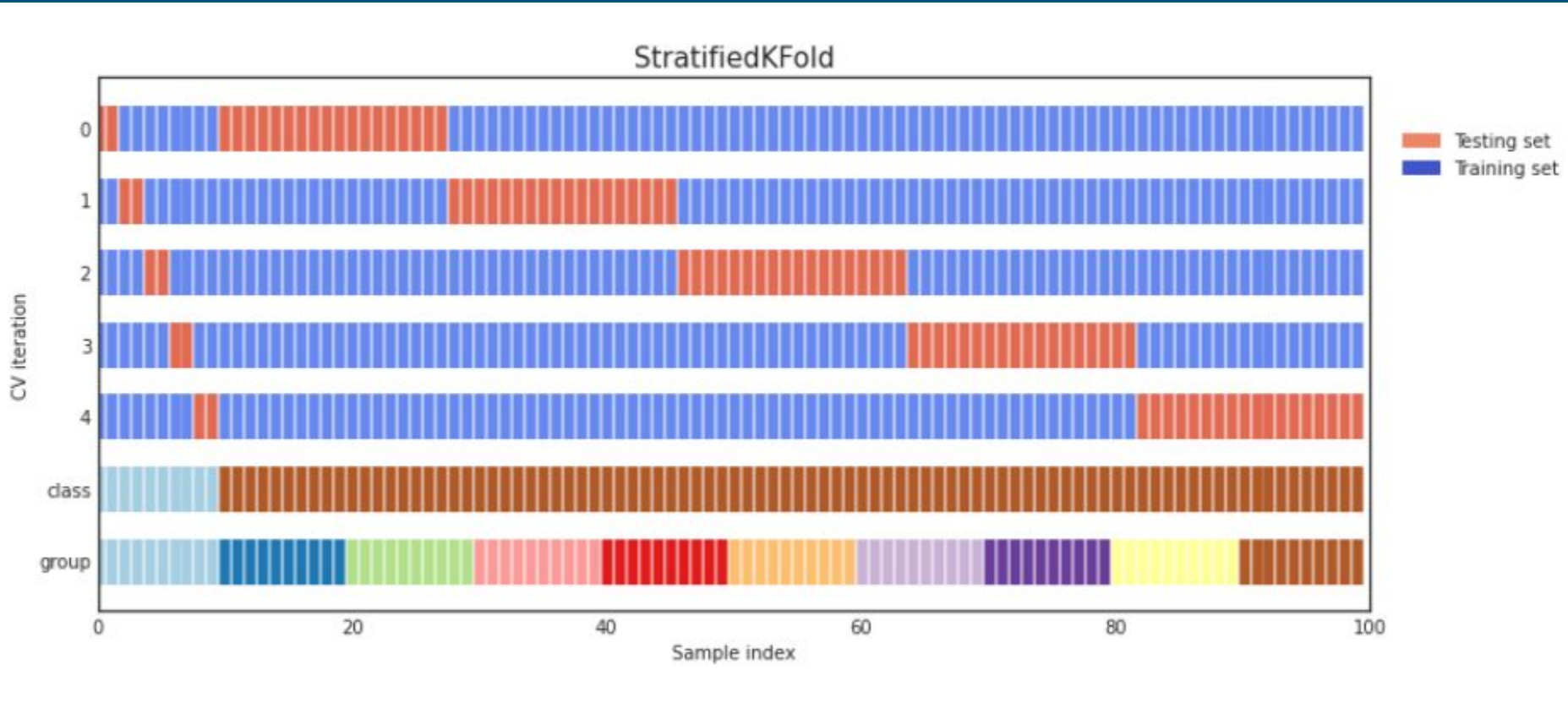
Thought Experiment

- What if you have a cancer dataset and 950 rows have negative values for the target and 50 rows have positive values for the target.
- If you did KFold Cross validation, how many positive target rows would be in some of the folds?

Thought Experiment

- Soln- what if we changed our cross validation so it had the same amount of positive and negative target values in each fold?
- This is called StratifiedKFold Cross Validation

Statified KFold CV (ignore group row)



Note

- StratifiedKFold only makes sense if you are doing classification
- Just KFold for Regression

Any questions?

- Let’s look at code
- Let’s do this with the Taylor Swift Dataset
- Go through how to do cross validation and print out the indices and the train/test sets

Other Applications for Cross Validation

- This is what we will do next class
- Using it to validate different feature selection methods
  - Wrapper methods like RFE
- Using it to explore hyperparameters for your models

If there is time

- Talk about training set vs validation set vs testing set
- Start talking about how different models work
  - Decision trees- search tree/if statements
  - Support vector machines- boundary predictor
  - Neural net- minimizing loss from gradient