

Introduction to Machine Learning

Day 8: The Full Pipeline

Joanna Bieri DATA301

Get out your handwritten notes!

Doing It By Hand

ASK

Last week you filled in the gaps,
made the dummy columns, and scaled,
all by hand.

What could go wrong?

Hint. Which rows did your median come from?

Doing It By Hand

ANSWER

- **Two things.** You can use rows you were not supposed to, and you have to redo every step by hand on every new set of data.
- Over everybody the median age is 28. Over the training rows only it is 29 instead.
- A year of age will not ruin the model. That is why this mistake survives: nothing announces it.

Counting Columns

PREDICT

column	what is in it
age	22, 38, 26, ...
sex	male, female
port	S, C, Q

One-hot encode the text and leave the number alone. How many columns come out?

Hint. One new column per category. The number column stays one column.

Counting Columns

ANSWER

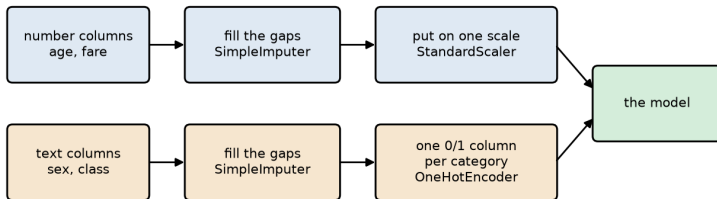
age	-> 1 column
sex	-> 2 columns (male, female)
port	-> 3 columns (S, C, Q)
total 6	

Our real table today turns 9 columns into 17 the same way.

Two Paths

ASK

One Pipeline: every step below is learned on training rows only



Why do the numbers and the text need different treatment?

Hint. What would it mean to fill in a missing port with a median?

Two Paths

ANSWER

- A missing **age** gets a number, the median. A missing **port** gets a category, the most common one.
- Numbers go on one scale. Text has to become numbers at all.
- ColumnTransformer sends each group down its own path, and they rejoin at the model.

A Passenger With No Age

ASK

A passenger has no age recorded.

The pipeline does not complain.

What did it put there?

Hint. And where did that number come from?

A Passenger With No Age

ANSWER

- The median age, learned by SimpleImputer from the **training rows only**.
- Then the scaler shifts it, using a mean and a spread also learned from the training rows.
- The row reaches the model as an ordinary number, and nothing was computed from data we are not allowed to use.

A Port It Has Never Seen

ASK

Next year a list arrives with a port code
your training data never had.

Does your model predict, or crash?

Hint. There is a flag in the encoder that decides this.

A Port It Has Never Seen

ANSWER

```
OneHotEncoder(handle_unknown="ignore")
```

With the flag, the unknown port becomes all zeros and the model still answers. Without it, the encoder raises an error and your program stops. In a live system that is the difference between a prediction and a phone call at 3am.

Trying Another Model

ASK

You want to try boosting instead
of logistic regression.

How much has to change?

Hint. The preprocessing does not know which model is on the end.

Trying Another Model

ANSWER

```
("model", XGBClassifier())
```

One line. That is the payoff for putting the preparation in a pipeline: comparing models becomes a fair comparison, because they all get identical data.

Counting Fits

PREDICT

```
{"model__n_estimators": [100, 300],  
 "model__max_depth": [None, 5, 10],  
 "model__min_samples_leaf": [1, 5]}
```

```
GridSearchCV(pipe, grid, cv=5)
```

How many times does this fit a model?

Hint. Multiply the choices together, then multiply by the folds.

Counting Fits

ANSWER

$2 * 3 * 2 = 12$ combinations

$12 * 5$ folds = 60 fits

Sixty fits, which takes a few seconds. Add a setting with four choices and the count jumps to 240 fits, which is where randomized search comes in.

Two Underscores

ASK

In the search you write
model__n_estimators,
not n_estimators.

Why?

Hint. There are several steps in the pipeline. Which one owns that setting?

Two Underscores

ANSWER

- The name says which step: the `n_estimators` belonging to the step called `model`.
- It nests as deep as you like. `prep__num__fill__strategy` reaches the imputer inside the number path.
- That last one would let you search over whether to fill with the median or the mean.

After The Search

ASK

The search used training and validation
rows together, in its folds.

Can we score the winner
on the validation set?

Hint. Where were those rows during the search?

After The Search

ANSWER

- **No.** Those rows were inside the search, so the winner has already seen them.
- The folds did the validating. That is why we did not hold a separate validation set out.
- The next honest number comes from the test set, which we open one time only.

Reading The Result

PREDICT

```
[[127  10]  
 [ 34  52]]
```

accuracy 0.803

86 passengers survived. What is this model actually good at?

Hint. Read the bottom row before you read the accuracy.

Reading The Result

ANSWER

- It finds **52 of the 86** survivors and misses 34 of them.
- It is much better at predicting who died than who lived, and the accuracy hides that.
- The asymmetry is the sentence that belongs in the report.

Sending It To Somebody

ASK

You email a colleague your trained model
and the raw passenger list.

What goes wrong?

Hint. What does the model expect its input to look like?

Sending It To Somebody

ANSWER

- The model wants 17 prepared columns and the raw list has 9 of mixed types.
- Your colleague also has no way to know your medians or your category lists.
- Save the **pipeline** instead of the model, so the preparation travels along too.

Questions?

Questions?