Building Data Narratives: An End-to-End Machine Learning Practicum

DataPhilly Workshops Fall 2020

Paul J Kowalczyk

Senior Data Scientist / Solvay

2020-09-17 13:34:46

Happy to hear from you . . .

- pauljkowalczyk@gmail.com
- linkedin.com/in/pauljkowalczyk/
- github.com/pjkowalczyk

Data Science Bucket List

✓Attend a DataPhilly meeting wearing pajamas.

In all sincerity . . .

. . . I hope that you, your family, and your friends are well.

And I thank you for taking time today to attend this practicum.

Cuppa Calculations

RECOMMENDED BREW TIMES

Tea Type	Brew Time
Black	4 mins
Chai	5 mins
Green	2 mins
Herbal	4 mins
Red	4 mins
Oolong	3 mins
White	1 min
Cold Brewed Iced Tea	5 mins

https://twiningsusa.com/pages/how-to-brew-the-perfect-cup-of-tea

Practicum

Merriam-Webster: "a course of study . . . that involves the supervised practical application of previously studied theory"

from the Latin, neuter of *practicus*: practical

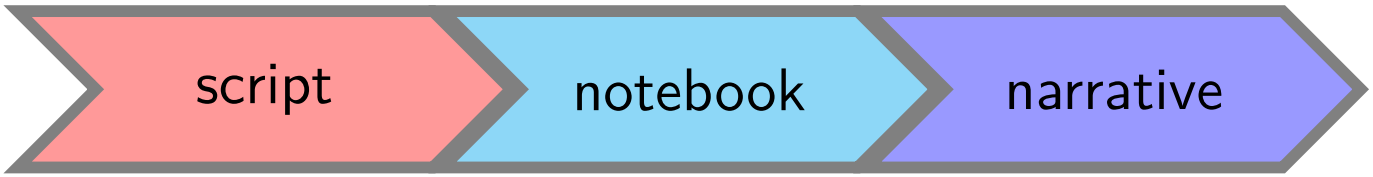
Reproducibility: Transparency

Reproducibility: Interoperability

(Purposefully) Serverless

- deliver narratives via *.html, *.pdf, *.doc, and *.pptx
- maintain interactivity

Workflow View - from a distance



- script: *.r, *.py
- notebook: *.Rmd, *.ipynb
- narrative: *.html, *.pdf, *.pptx, *.doc

The actual code is the same throughout. The difference is in how that code is **decorated**.

The Narratives' Audience

Creator

develops a computational solution to a problem; writes (& comments!) the scripts

Curator

uses the scripts:

- same script / different data
- repurposes the script
- expands the script
- ...

Consumer

exploits the actionable insights delivered in the narratives

These roles are not mutually exclusive.

Dataset Template

ID	**EndPt**	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	...	$\mathbf{x_n}$
ID_1	y_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	...	$x_{1,n}$
ID_2	y_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	...	$x_{2,n}$
ID_3	y_3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	...	$x_{3,n}$
...	...	...	...	...	...	...
ID_m	y_m	$x_{m,1}$	$x_{m,2}$	$x_{m,3}$	...	$x_{m,n}$

- **m** rows $\equiv$ samples
- **n** columns $\equiv$ features
- **EndPt** $\equiv$ dependent variable
- **x** $\equiv$ independent variable

To use a local dataset with the code from these exercises, label the column of identifiers **ID** (note capitalization), and label the column of dependent variables (**y**) **EndPt** (again, note capitalization).

Confusion Matrices

Form I

	Prediction	
True Label	**Negative**	**Positive**
Negative	TN	FP
Positive	FN	TP

Form II

	True Label	
Prediction	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

Form I: Zumel & Mount 'Practical Data Science with R'; scikit-learn

Form II: R::caret

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Kappa} = \frac{2*(TP*TN-FN*FP)}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{FP+TN}$$