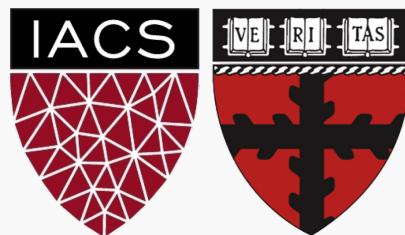


Lecture 16: Lingering Questions and A Basic Case Study

CS109A Introduction to Data Science
Pavlos Protopapas, Kevin Rader and Chris Tanner



Announcements

1. **HW4** is up
2. **Project** topic assignments
 - Released tomorrow (every group gets assigned a TF)
 - Every full group got their first choice
 - Those with custom topics will take a day or two more processing time.
3. No **Lecture** on Monday! No **Sections** Fri or Mon. No OHs on Mon.
4. Next **Advanced Section**: next week (Oct 14)



Lecture Outline

- A Little Review: some lingering questions
- Some HW feedback
- The Data Science Process: A Basic Case Study



Some Lingering Questions



What's the deal with the Intercept?

We've seen many forms of parametric models: linear regression, LASSO and ridge regression, and now logistic regression:

When all my X 's are zero, it's the predicted

$$\mu_Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{linear}$$
$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{logistic}$$

What does β_0 represent? Why do we need it? When do we not need it?

Modeling specifics: you almost always (>99% of the time) want to have an intercept (sometimes called the bias term) in your model. Why?

1) When y is already centered at zero
2) force the line to go through the origin

sklearn specifics: linear and logistic regression include an intercept by default (yay!). Polynomial features creates the bias term by default (boo). You don't want both, and if you are regularizing, you don't want to shrink it!

3) Binary predictors:
and $y \in \{0, 1\}$
don't want
a reference



What's the difference: Standardize vs. Normalize

What is the difference between Standardizing and Normalizing a variable?

- Normalizing means to bound your variable's observations between zero and one. Good when interpretations of “percentage of max value” makes sense.
- Standardizing means to re-center and re-scale your variable's observations to have mean zero and variance one. Good to put all of your variables on the same scale (have same weight) and to turn interpretations into “changes in terms of standard deviation.”

Warning: the term “normalize” gets incorrectly used all the time (online (like in sklearn), especially)!



What's the difference: Standardize vs. Normalize

What is the difference between Standardizing and Normalizing a variable?

- Normalizing means to bound your variable's observations between zero and one. Good when interpretations of “percentage of max value” makes sense.
- Standardizing means to re-center and re-scale your variable's observations to have mean zero and variance one. Good to put all of your variables on the same scale (have same weight) and to turn interpretations into “changes in terms of standard deviation.”

Warning: the term “normalize” gets incorrectly used all the time (online, especially)!



To Standardize, or Not to Standardize, that is the question...

When should you do each?

- Normalizing is only for improving interpretation (and dealing with numerically very large or small measures). Does not improve algorithms otherwise.
- Standardizing can be used for improving interpretation and should be used for specific algorithms. Which ones? Regularization and k-NN (so far!) *→ predictors*
- For best interpretations, do not change the scale from the original units.
*Note: you can standardize without assuming things to be [approximately] Normally distributed! It just makes the interpretation nice if they are Normally distributed.



(Train & Validation) & Test Splits

What is the proper use of Train, Validation, and Test Splits? What is the purpose?

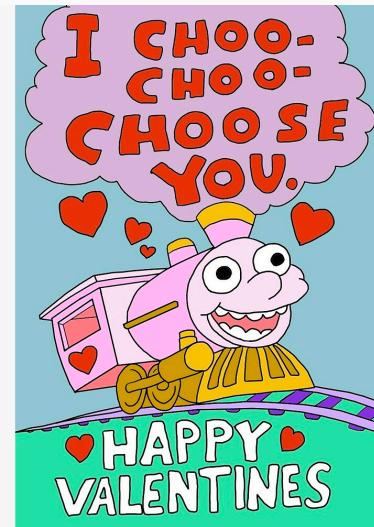
Purpose: to select and evaluate a **predictive** model. Not needed for interpretative models.

- The **test** set is used only for evaluating the error of the model (and used just once).
- The **(train & validation)** sets are used to estimate/fit and select between models, respectively.
- Since it is so important to choose your best predictive model, we often perform cross validation so that a single random validation set does not influence this decision.

Note: you should refit the model on the entire (Train & Validation) set after choosing the best model.



It's funny, because it's a train



A few other lingering thoughts:

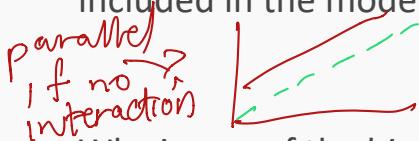
correlated X 's

it's impossible
to hold other
predictors
constant

What happens when multicollinearity is present? Why is multicollinearity bad?
Why is it not always bad? How can this be handled?

"controlling for confounders"

How does a multiple regression model for 2 predictors (one binary, the other quantitative) compare in interpretation when the interaction is included vs. not included in the model. How can we visualize this?



w/ interaction



in β -estimates

Why is one of the binary/dummy terms dropped out of the predictor set when modeling a categorical predictors?

if intercept and all dummies
 \Rightarrow perfect collinearity



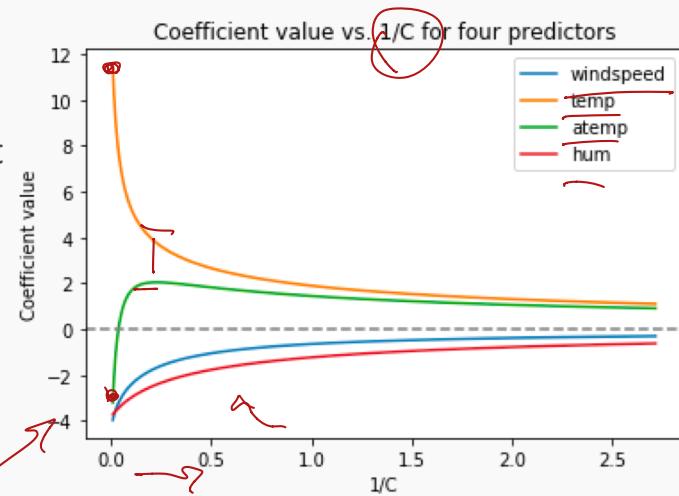
Some Regularizing Details

default prediction model
to be $\hat{y} = \bar{y}$ not $y = 0$

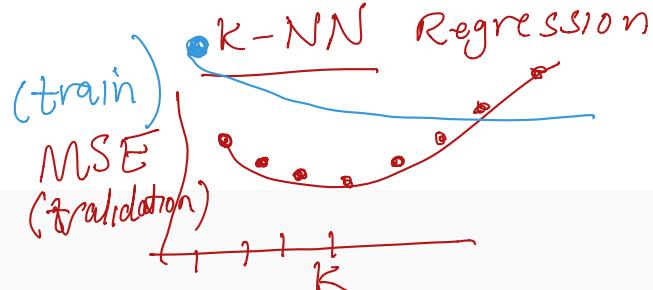
Why is the intercept term not penalized in regularized methods? What if it was penalized?

When is LASSO preferred over Ridge? What about the other way around?

How do regularization methods affect multicollinearity?



$\lambda \uparrow$
(more regularization)



Suggestions/Comments about HW

- Plotting your model's predictions (presumably on a scatterplot with real data) use np.linspace to create a "dummy x" and connect with lines and not points!
- Sorting data: don't do it! What could go wrong? Sorting results, do it!
- How do you "compare the performance" of several models (to select a best k) in k -NN or λ in LASSO/Ridge? plot metric vs. "hyperparameter"
- Bar plot vs. Histogram: what if we have integers? Discrete few categories \Rightarrow barplot
- When bootstrapping (or simulating) to compare models/approaches: sample once and fit all models you're comparing on this same sample (not a separate sample for each model). Why? \leftarrow better "comparison groups"
- Significant digits!!! Avoid extremes: 0.00001 and 9.2857104571934812 are not the best choices. We are not super critical, but 3-5 sigdigs are reasonable.
- Be sure to re-run the entire notebook and label your output (don't just print out numbers without reference): we'll start taking off points in HW4.

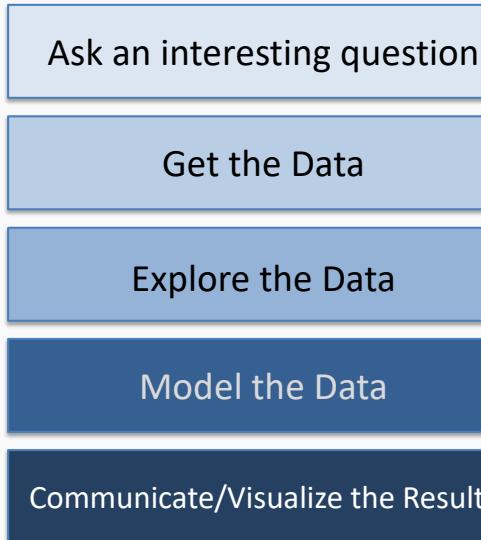


The Data Science Process: A 'Case Study'

Important for
your project,



The Data Science Process



Start here!

I started with some interesting questions. I wanted to address:

- **The wage gap.** There is actually lot of publicly available salary data out on the web. But the data do not include sex/gender/race/ethnicity, and are typically volunteered.
- **The effect of COVID-19 on various things: mental health, media consumption, etc.** Publicly available data is very spotty. Or not raw enough to be useful. Or very expensive (see, Nielsen cost).
- I browsed Kaggle Competitions for interesting questions they have. Nothing tickled my fancy.

not linear

So I called an audible...



Do you like...stuff?

Ask an interesting question

Get the Data

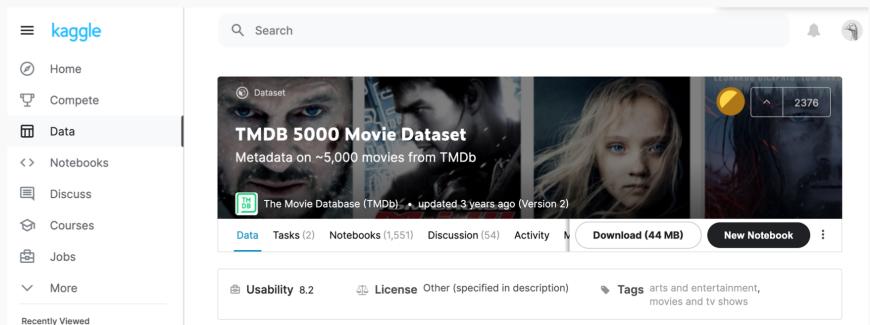
Explore the Data

Model the Data

Communicate/Visualize the Results

I like movies. Let's start our process by considering a data set from Kaggle:

<https://www.kaggle.com/tmdb/tmdb-movie-metadata>



What should we do first?



First Glimpse at the Data

Explore the Data

```
print(movies.dtypes)
```

<u>budget</u>	int64
<u>genres</u>	object
<u>homepage</u>	object
<u>id</u>	int64
<u>keywords</u>	object
original_language	object
original_title	object
overview	object
popularity	float64
production_companies	object
production_countries	object
<u>release_date</u>	object
revenue	int64
<u>runtime</u>	float64
spoken_languages	object
status	object
tagline	object
title	object
vote_average	float64
vote_count	int64
<u>dtype: object</u>	

```
movies = pd.read_csv('data/tmdb_5000_movies.csv')
credits = pd.read_csv('data/tmdb_5000_credits.csv')
```

```
movies.head()
```

budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_co
0 237000000	[{"id": 28, "name": "Action"}, {"id": 12, "name": "..."}]	http://www.avatarmovie.com/	19995	{"id": 1463, "name": "culture clash", "id": ...}	en	Avatar	In the 22nd century, a paraplegic Marine is d...	150.437577	["name": "Film Partn...

1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Action"}]	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "name": "na..."}]	en	Pirates of the Caribbean: At World's End	Captain Barbosa, long believed to be dead, ha...	139.082615	["name": "W... Pictures", "id": 1000000000]
		[{"id": 28, "name": "Romantic"}]			[{"id": 28, "name": "Romantic"}]			A cryptic message		

2	245000000	{"name": "Action", "id": 12, "name...": "Action"}, http://www.sonypictures.com/movies/spectre/ 206647	470, "me...": "spy", "id": 818, "name...": "name..."}	en	Spectre	meets Bond's past	107.376788	["name": "Pictures"]
---	-----------	---	---	----	---------	-------------------	------------	----------------------

```
  credits.head(
```

movie_id	title	cast	crew
0	1995	Avatar	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "de...", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "...", "de...

What are some necessary cleaning and wrangling tasks? What are some EDAs we should perform?

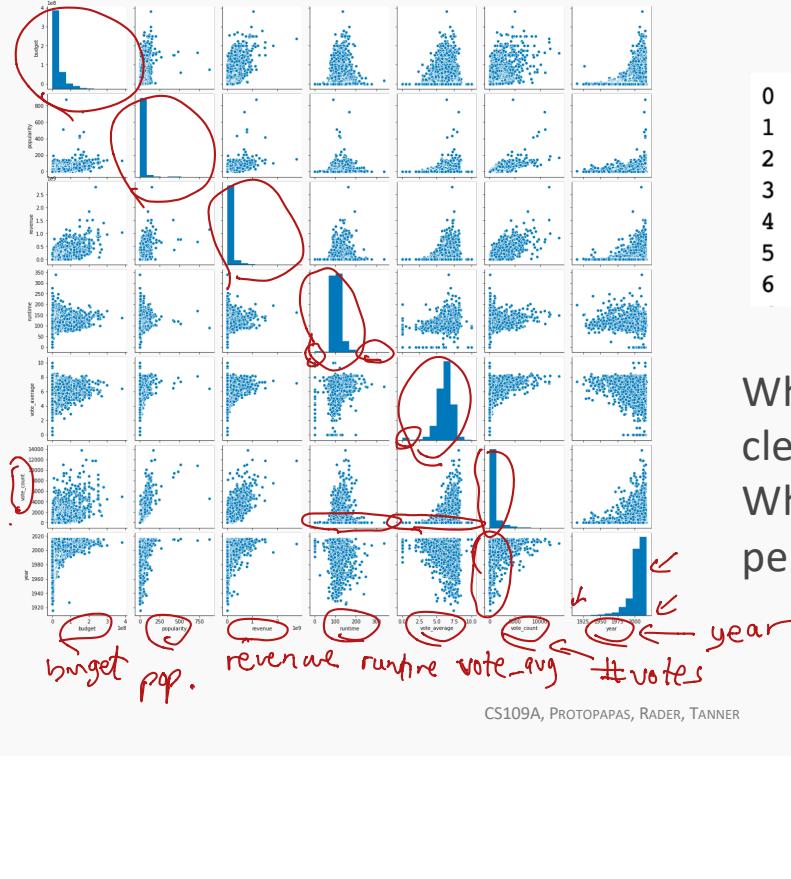


Three EDAs we should perform: correlation matrices, tables

Summaries {
 descriptive } distributions (barplots, histograms)
Visuals {
 univariate ← distributions
 bivariate ← scatterplots, groupby, side-by-side boxplots
 conditional barplots... .

First Glimpse at the Data

Explore the Data



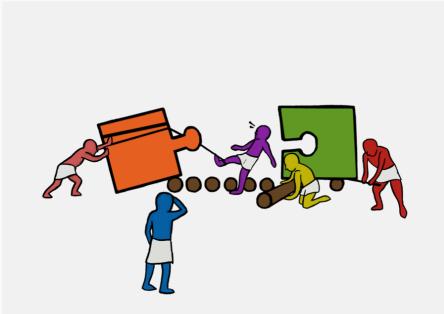
0 budget
1 popularity
2 revenue
3 runtime
4 vote_average
5 vote_count
6 year

budget
popularity
revenue
runtime
vote_average
vote_count
year

wrangle
from
release_data

What are some necessary
cleaning and wrangling tasks?
What are some EDAs we should
perform?





Breakout #1 Tasks (15-20min):

1. Someone share (the person who resides closest to the Bahamas...thanks Columbus). Someone different will share in the next breakout.
2. Explore the data (some of that is done with you with code). Please do a little more exploration.
3. Come up with an interesting question or two you can answer with this data set. Come up with a question or two that can be answered with supplemental data:
 - start with ideal, and then get more practical based on what is likely available.

Any interesting questions?

Ask an interesting question

What are some interesting questions you came up with for this data set?

Get the Data

Explore the Data

What are some interesting questions using supplemental data?

Model the Data

Communicate/Visualize the Results



My questions:

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

I came up with simple questions (and pretty straight-forward one for now):

#1: How is revenue associated with budget?

#2: How has this evolved over time?

#3: What other factors relate to revenue of a movie?



Back to exploring:

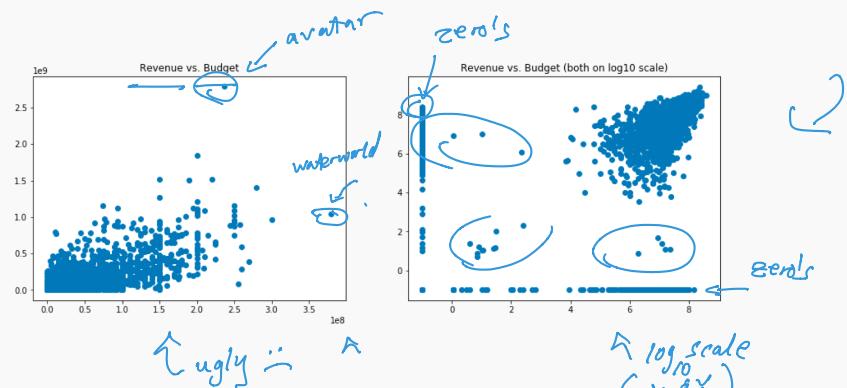
Ask an interesting question

Get the Data

Explore the Data

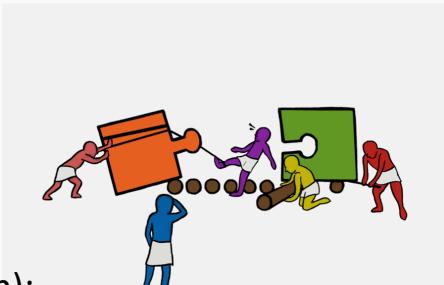
Model the Data

Communicate/Visualize the Results



Which is more appropriate for linear modeling
(or any regression modeling, really)? What
issues arise? What should we do about it?





Breakout #2 Tasks (20+min):

1. Someone else share and **take notes** (who resides furthest from the [Bahamas](#))
2. Solidify your question(s) of interest.

3. Determine the next tasks:

- What other data do you need? How will this data be collected and combined?
- What data cleaning and wrangling tasks are needed?
- What other EDA is necessary? What visuals should be included?
- What is a goal for a first baseline model (Key: should be interpretable)?

Be sure to include the class of model and the variables involved.

- What is a reasonable goal for a final model and product?

4. Determine how long each task should take.

*communicating
model results.*

5. Assign next tasks to group members. Do not actual perform these tasks!



Scoping and Redefining the Problem

How did that go?

What is the biggest concern you have at this point for your questions?

Can these data answer your questions?

What supplemental data would be required?

What idealistic questions did you shy away from?

What models are you considering?



* Custom Projects: please submit your proposal
* Check your Canvas Group for your project!

Scoping and Redefining your Project

* Look for a message
from assigned TF

This lecture is relevant for your group project:

Milestone 2 (due next week on Friday) is to scope and redefine your project.

You should converse with your group members to:

(a) Redefine the problem, if appropriate

- For example: maybe you want to not use the Boston Crime data set, but instead use the FBI's data for white collar crime).

(b) Scope your project

- What are the next tasks? Who will perform what ?
- What is a reasonable goal for a baseline model?
- What are both ideal and reasonable goals for a final model and product?)

*feel free to consult with your assigned TF in a message or two.



My question:

What models should be considered for these questions?

Ask an interesting question

#1: How is revenue associated with budget?

Get the Data

#2: How has this evolved over time?

Model the Data

#3: What other factors relate to revenue of a movie?

Communicate/Visualize the Results

How could these models be communicated?

