God, Beer, and Sports

Predicting sports fans from OKCupid survey data

Adam Flugel - Metis Chicago, Winter 2018

Client - Boiz w/ Ballz

- Sports marketing company
- Seeking to better understand audience
- Targeting campaigns

Client - Boiz w/ Ballz

- Sports marketing company
- Seeking to better understand audience
- Targeting campaigns

Objective

What kind of people tend to enjoy watching sports?

- What personal questions are predictive?
- Create app to test individuals for potential sports interest

The Tools

SQLAlchemy

amazon web services™ | EC2

Flask

web development, one drop at a time

fancyimpute 0.3.1

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

NumPy

SHARE

SHARE 1933

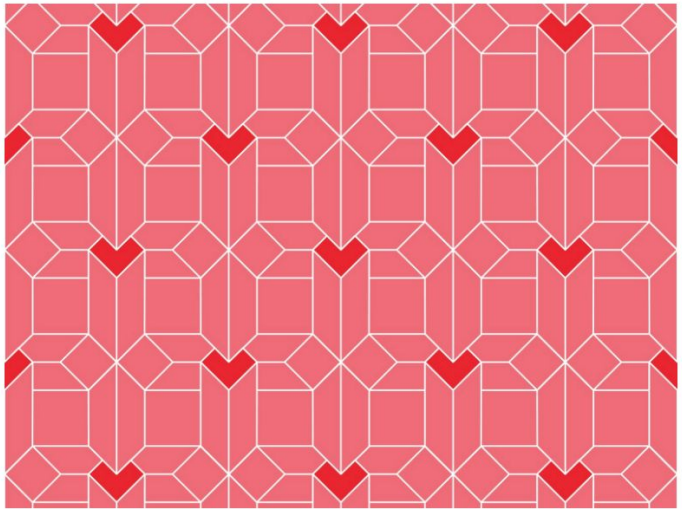
TWEET

COMMENT

EMAIL

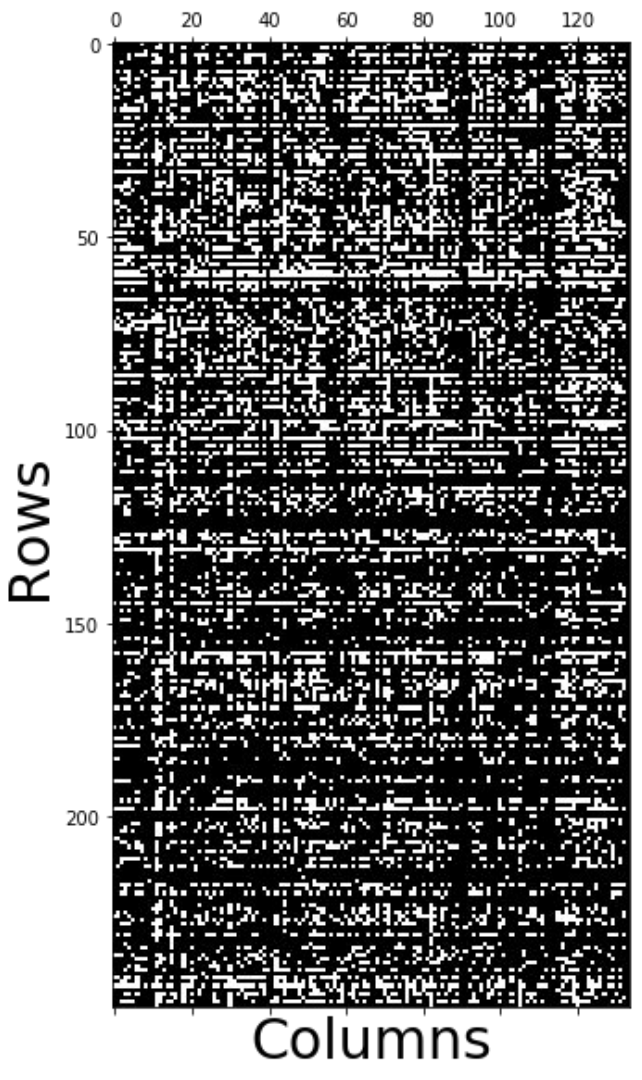
MICHAEL ZIMMER OPINION 05.14.16 07:00 AM

OKCUPID STUDY REVEALS THE PERILS OF BIG-DATA SCIENCE



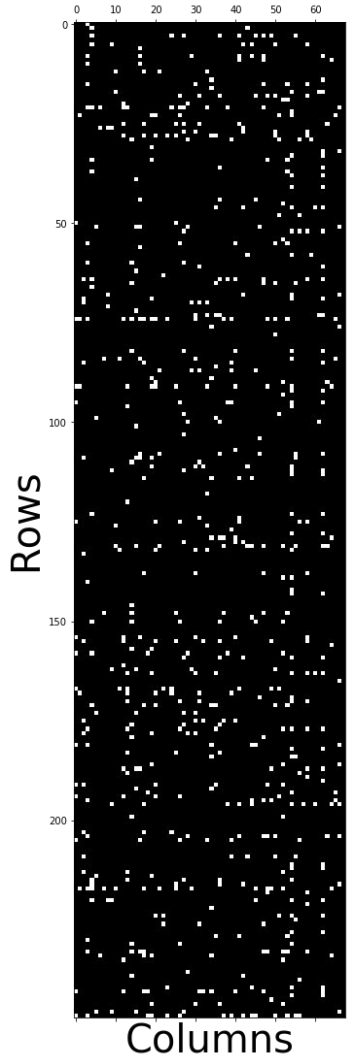
OKCupid User Data

- ♥ 68,371 users
- ♥ 2621 questions
- ♥ Kept ‘yes/no’ questions



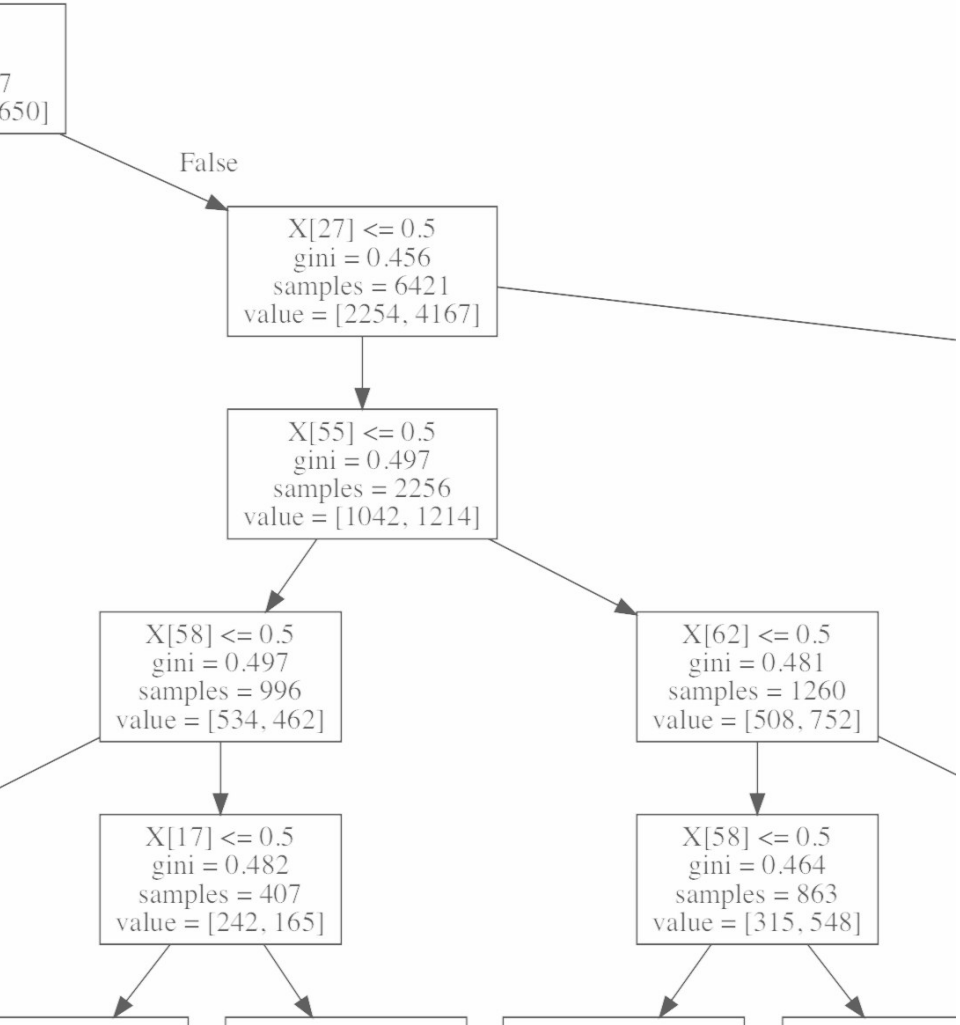
Missing Data

- Systematic!
- Dropped
  - rows missing >20%
  - Cols missing >15%



Final Dataset

- ♥ 21,237 users
- ♥ 67 questions
- ♥ MICE to fill missing values



Features 1 (decision tree):

Should burning your country's flag be illegal?

Do you like the taste of beer?

Do you think women have an obligation to keep their legs shaved?

Are you Christian?

Were you picked on a lot in school?

Do you believe in God?

Are you either vegetarian or vegan?

If you like someone a lot, do you usually ask them out?

`sklearn.feature_selection.SelectKBest`

Features 2 (Chi2):

Would it be useful and ethical to clone the best and brightest of our species, for the common good?

Do you think women have an obligation to keep their legs shaved?

Should burning your country's flag be illegal?
Do you believe in God?

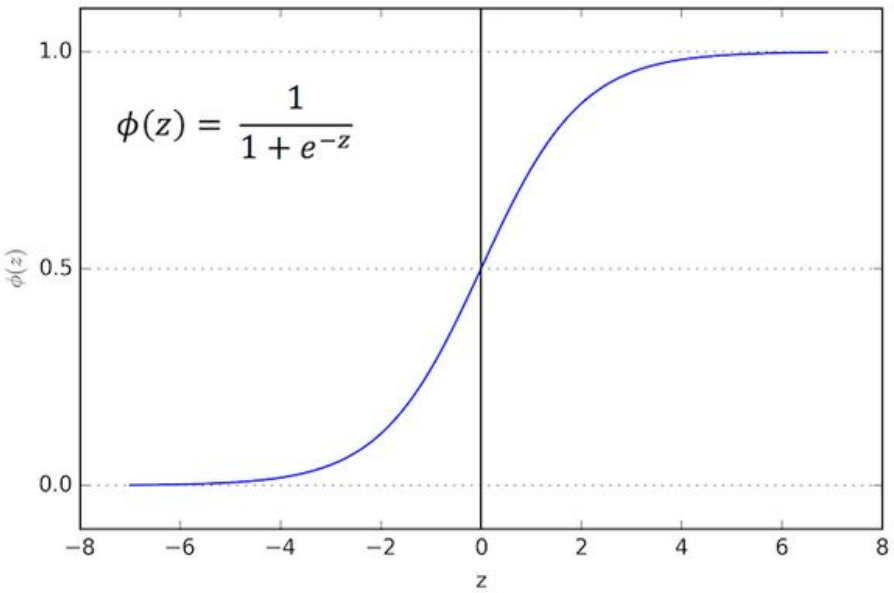
Are you Christian?

Are you either vegetarian or vegan?

Were you picked on a lot in school?

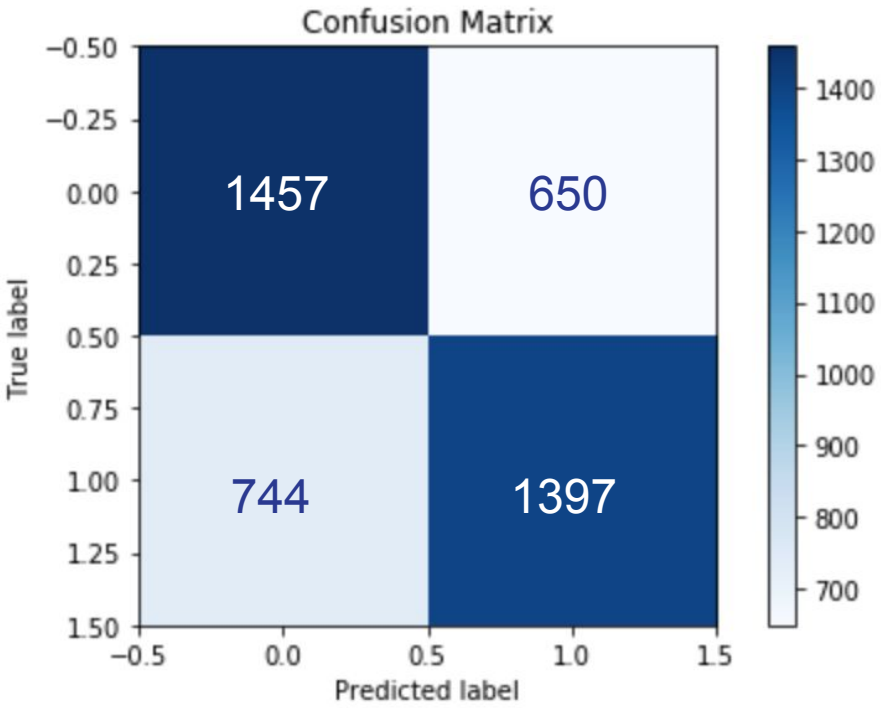
Are you an aspiring actor/artist/writer or other creative type?

\+ 8 more



Best Model

- ♥ Logistic Regression
- ♥ C = 0.1
- ♥ **All features used**

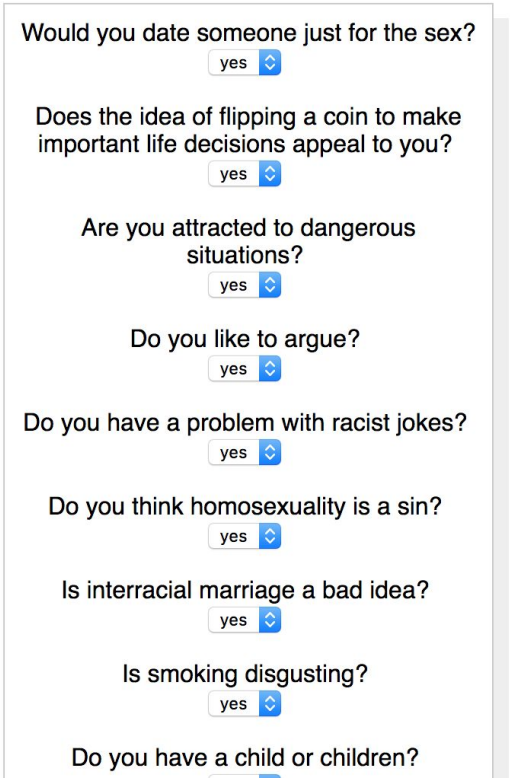


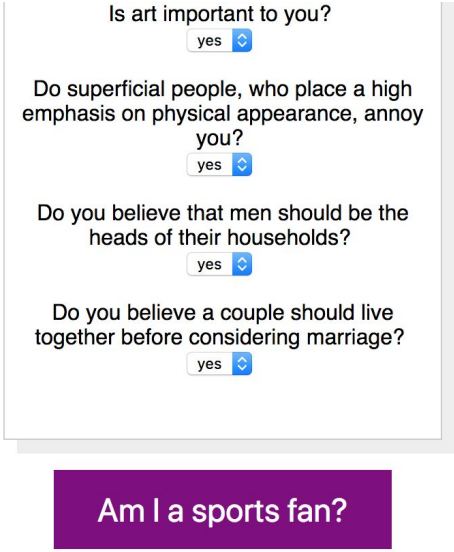
Scores

- ♥ Accuracy = 67%
- ♥ Precision = 66%
- ♥ Recall = 69%
- ♥ Log loss = 0.62

Flask App

Answer the questions and I'll tell you if you're a sports fan.





Am I a sports fan?

You love those rough and tumble games so much!

Actionable Insights?

What does this data tell us?

- ♥ “Good ol’ boys” love sports

Actionable Insights?

What does this data tell us?

- ♥ “Good ol’ boys” love sports

BUT

Actionable Insights?

What does this data tell us?

- “Good ol’ boys” love sports

BUT

- Not all sports lovers are “good ol’ boys”

Going Forward...

Further exploration

- Work with features besides yes/no/maybe
- Use app data/feedback to pinpoint where model struggles

Appendix

Other analysis

Adam Flugel - Metis Chicago, Winter 2018

Comparing various dataset transformations

	q17	q55	q56	q57	q65	q70	q71	q87	q105	q114	...
no_impute	-0.003195	0.019432	-0.021726	0.050637	0.080305	-0.044098	-0.005946	0.006533	-0.057041	0.018929	...
impute_continuous	-0.002521	0.019156	-0.021655	0.050165	0.078986	-0.045048	-0.006499	0.006901	-0.056370	0.018986	...
impute_binary	-0.002517	0.018820	-0.022408	0.050013	0.084681	-0.043375	-0.005611	0.007868	-0.056752	0.018584	...

Difference between mean responses of target groups (sports fans vs. not) per column

Decision tree feature importance

```
[('q175', 0.3566131071005708 7),
 ('q1112', 0.1991532247841227 3),
 ('q134', 0.151076042606696 55),
 ('q156913', 0.079489700264260 62),
 ('q313640', 0.073645354383816 1),
 ('q210', 0.028874345193184593),
 ('q179268', 0.022816429499781875),
 ('q308', 0.015782603204391798),
 ('q1454', 0.013487063473920749),
 ('q325', 0.013097023391952236),
 ('q158', 0.011716543549022315),
 ('q403', 0.007495690702000676),
 ('q70', 0.006744995074152282 6),
 ('q80041', 0.006318663492679137 5),
 ('q784', 0.004614974105938085)]
```

Cutoff