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

The Tools

SQLAlchemy

amazon web services™ | EC2

fancyimpute 0.3.1

pandas

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

NumPy

The Tools

SQLAlchemy

amazon web services™ | EC2

fancyimpute 0.3.1

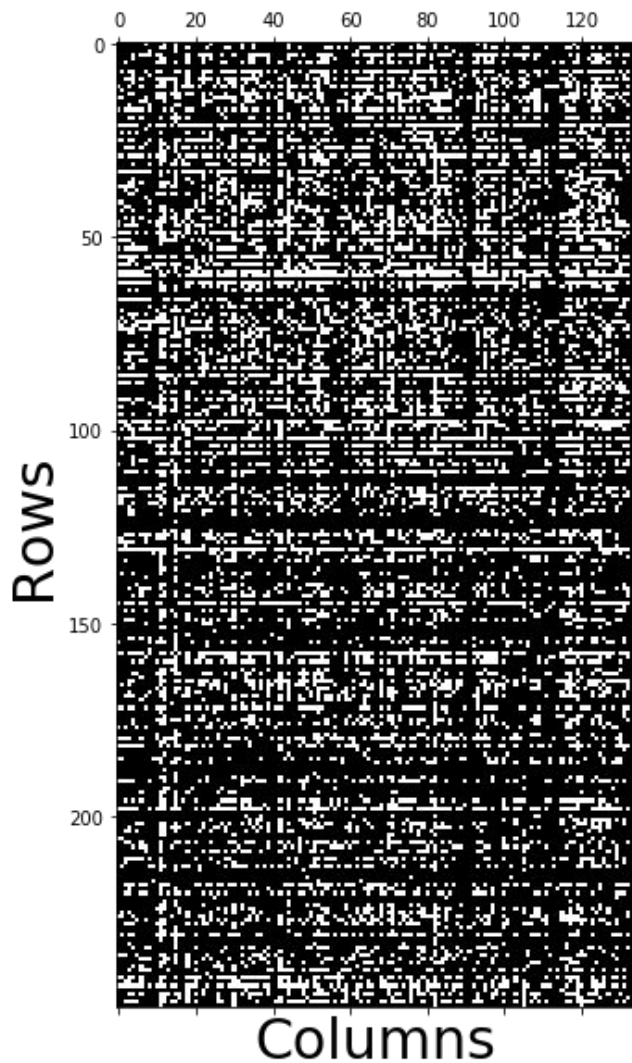pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$ NumPy
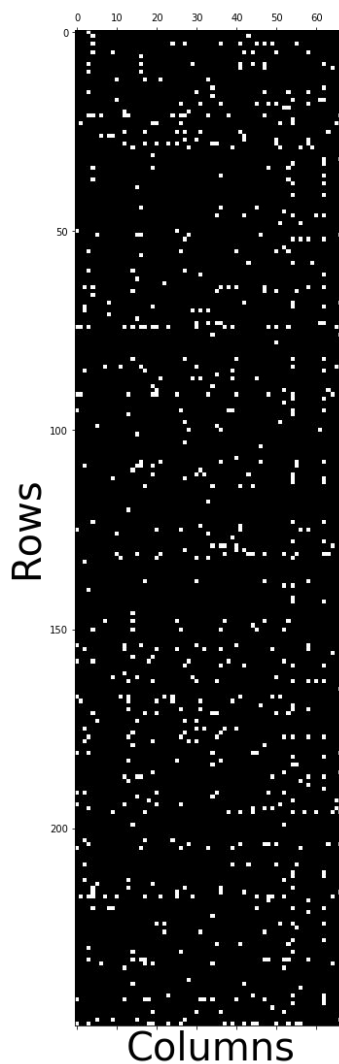
scikit learn

The Tools
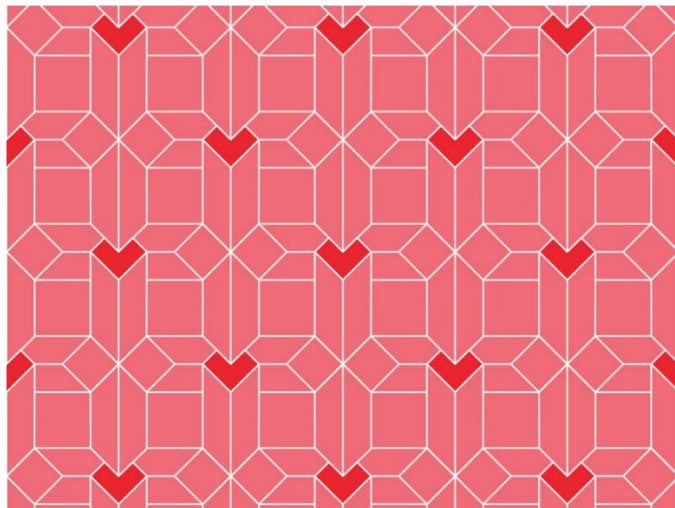
OKCupid User Data

♥ 68,371 users

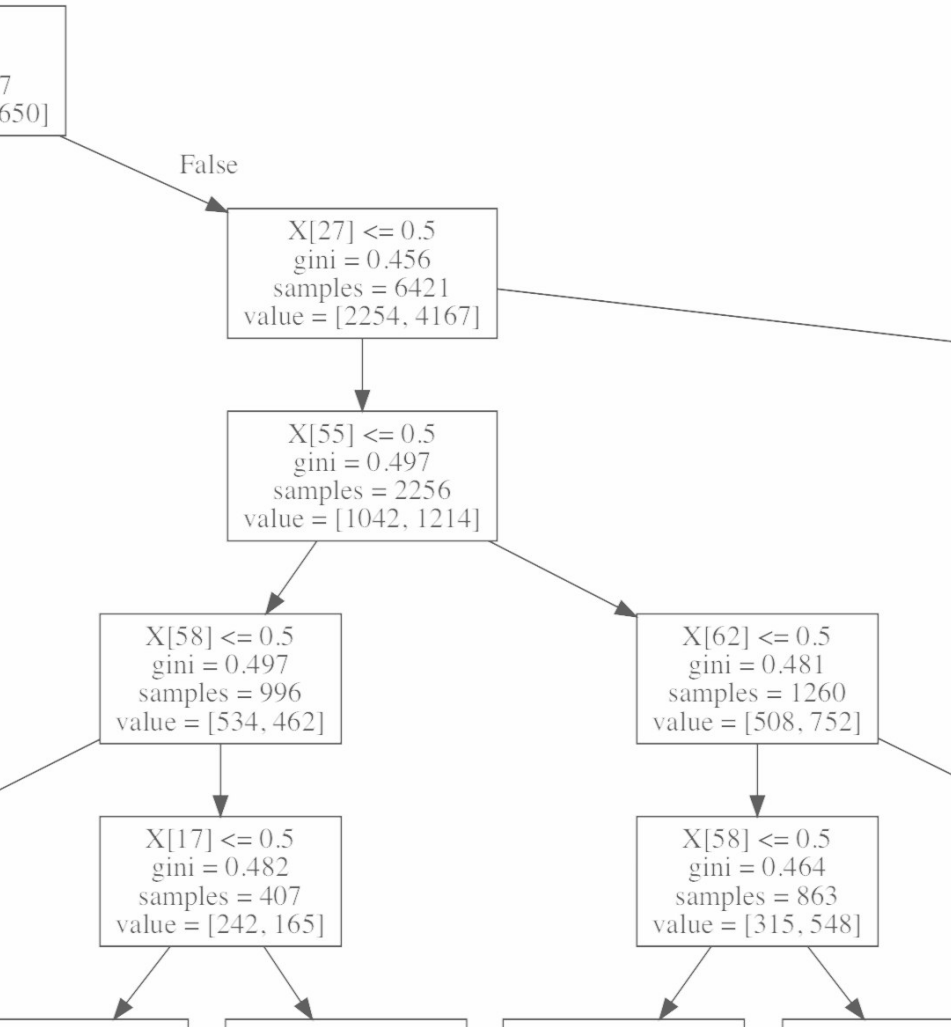♥ 2621 questions

♥ 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?

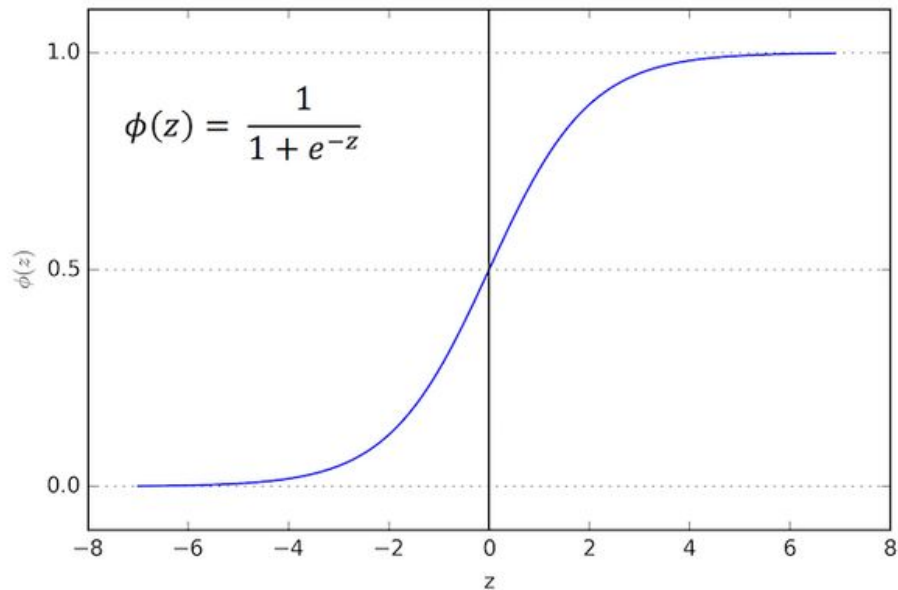**Do you believe in God?**

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?

+   7 more

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

Best Model

♥ Logistic Regression

♥ C = 0.1

♥ **All features used**

Confusion Matrix

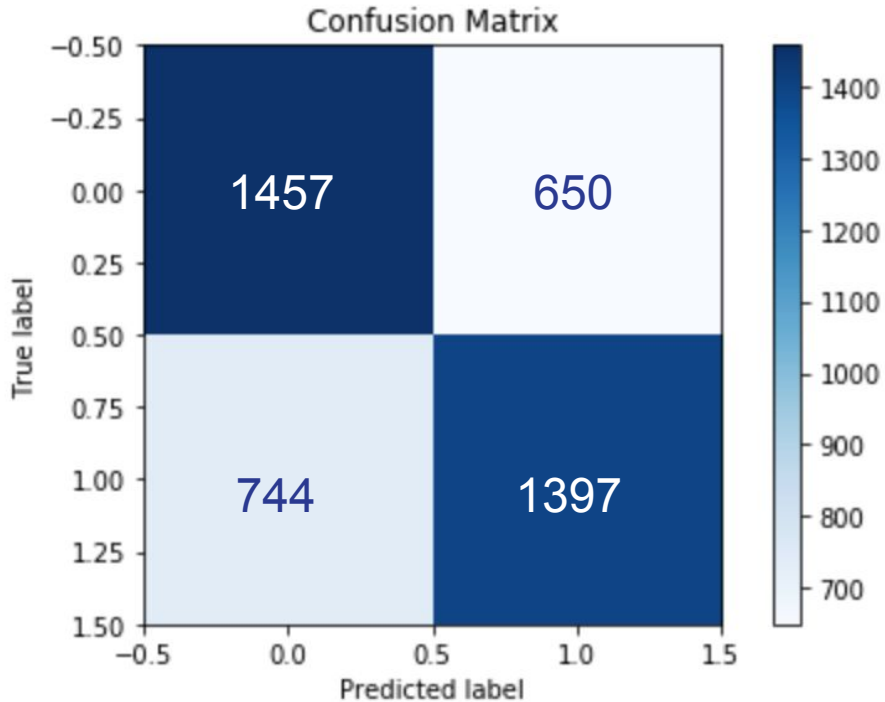|	Predicted 0	Predicted 1
True 0	1457	650
True 1	744	1397

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.

Would you date someone just for the sex?
yes

Does the idea of flipping a coin to make important life decisions appeal to you?
yes

Are you attracted to dangerous situations?
yes

Do you like to argue?
yes

Do you have a problem with racist jokes?
yes

Do you think homosexuality is a sin?
yes

Is interracial marriage a bad idea?
yes

Is smoking disgusting?
yes

Do you have a child or children?

Is art important to you?
yes

Do superficial people, who place a high emphasis on physical appearance, annoy you?
yes

Do you believe that men should be the heads of their households?
yes

Do you believe a couple should live together before considering marriage?
yes

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', 0.35661310710057087),
 ('q1112', 0.19915322478412273),
 ('q134', 0.15107604260669655),
 ('q156913', 0.07948970026426062),
 ('q313640', 0.0736453543838161),
 ('q210', 0.02887345193184593),
 ('q179268', 0.022816429499781875),
 ('q308', 0.015782603204391798),
 ('q1454', 0.013487063473920749),
 ('q325', 0.013097023391952236),
 ('q158', 0.011716543549022315),
 ('q403', 0.0074956907020000676),
 ('q70', 0.0067449950741522826),
 ('q80041', 0.0063186634926791375),
 ('q784', 0.00461497410593808)]
```

Cutoff