

# Predicting Gender from OKCupid Profiles

## Using ensemble methods

Isaac Backus

### Feature Generation

- Drop short response answers
- Convert to multiple choice to binary:

pets	Has dogs	Likes dogs	Has cats	Likes cats
“has dogs and likes cats”	1	1	0	1

- This generates ~215 features
- Also created a smaller set by condensing education, astrological sign, and languages (~100 features)

### Dataset

- 59946 OKCupid profiles from 2011
- San Francisco only
- Gender **m/f** only
- 40% female, 60% male
- 10 short response questions and
- 20 multiple choice or number responses

### Replace Missing Data

10% of questions not answered

Techniques:

- 1) *Mean imputation* – replace missing values with feature mean
- 2) *Random selection* – replace with value randomly chosen from dataset
- 3) *Regression model*:
  - Do mean imputation
  - Do linear regression on all features
  - Replace with predicted values
  - Fails! (outliers)

### Train models

- Gaussian Model
- K-means (256 clusters)
- L2 Boosted Ridge Regression
- Ridge Regression
- Logistic Regression
- Linear Random Map w/ logistic
- RBF Kernel Logistic
- Random Forest (sklearn)

### Ensemble Predictions (stacking)

Each model returns a number  $f(x)$  in  $[-1, 1]$   
 $\text{sign}(f(x)) = \text{class}$

- $|f(x)|$  = confidence of the prediction
- Find  $a_m$  to minimize 0/1 loss for

$$F(x) = \sum_m a_m f_m(x)$$

The predicted class given by  $\text{sign}(F(x))$

## Results

### 0/1 Losses

Method	Original Set		Pared features	
	Training	Test	Training	Test
Gaussian Model	0.316	0.316	0.246	0.255
K-means (256 clusters)	0.301	0.307	0.291	0.311
L2 Boosted Ridge Regression	0.113	0.119	0.115	0.123
Ridge Regression	0.113	0.119	0.115	0.123
Logistic Regression	0.112	0.119	0.114	0.121
Linear Random Map	0.143	0.156	0.124	0.133
Random Forest (sklearn)	0.116	0.128	0.114	0.128
RBF Kernel Logistic	0.171	0.185	0.160	0.162
<b>Ensemble</b>	<b>0.106</b>	<b>0.116</b>	<b>0.106</b>	<b>0.118</b>

### Strongest predictors

Male	w	Female	w
height	0.279	body_type_curvy	0.071
body_type_athletic	0.042	ethnicity_white	0.040
job_computer / hardware / software	0.035	body_type_full figured	0.035
orientation_gay	0.026	orientation_bisexual	0.031
ethnicity_hispanic / latin	0.024	has_cats	0.029
job_science / tech / engineering	0.022	job_medicine / health	0.022

### Weakest predictors

Lowest weights	w
speaks_latvian	1.10E-09
speaks_serbian	1.16E-05
education_ph.d program	2.13E-05
sign_taurus	5.53E-05
speaks_lisp	7.66E-05

