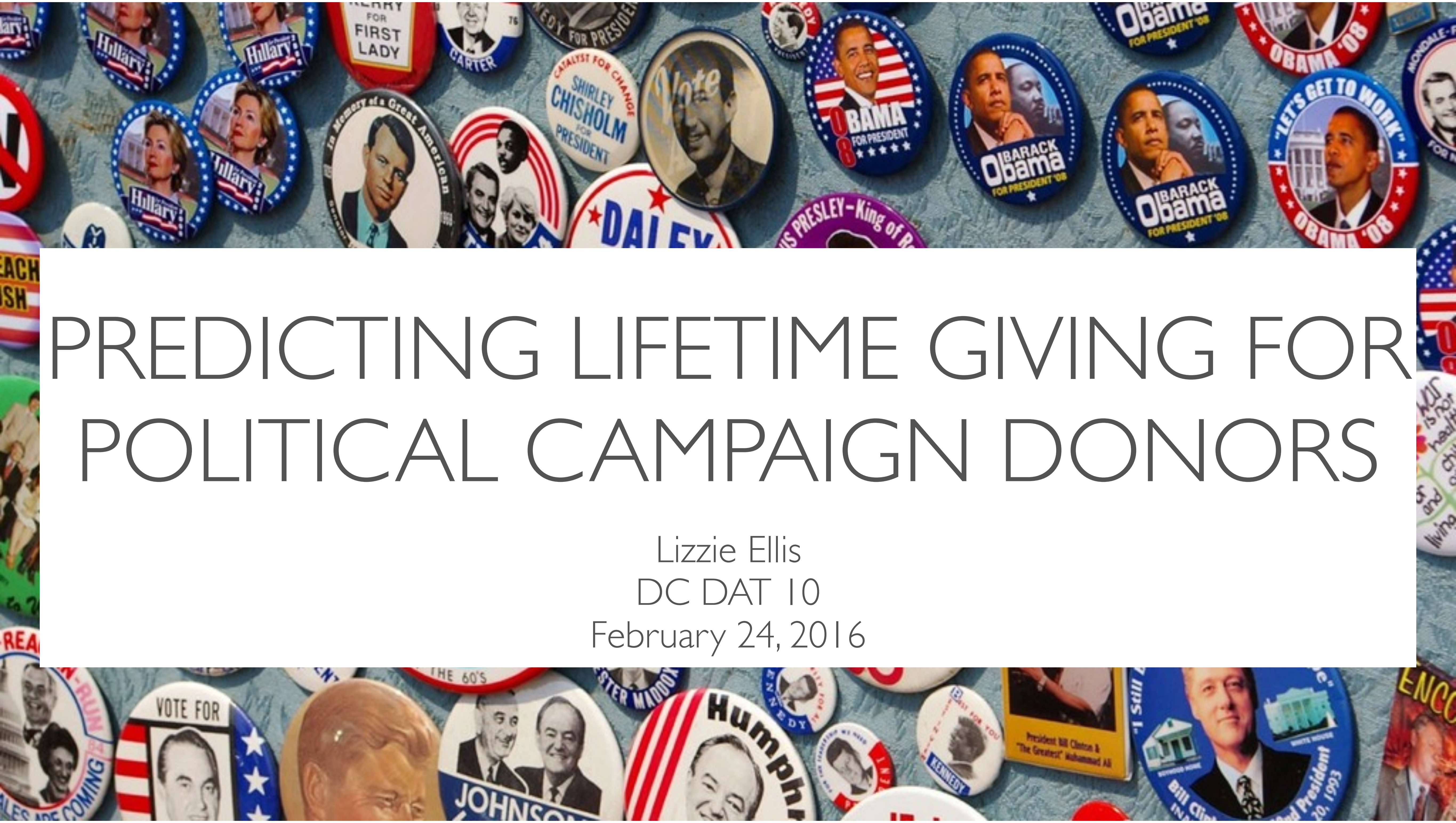


PREDICTING LIFETIME GIVING FOR POLITICAL CAMPAIGN DONORS

Lizzie Ellis
DC DAT 10
February 24, 2016

THE GOAL:

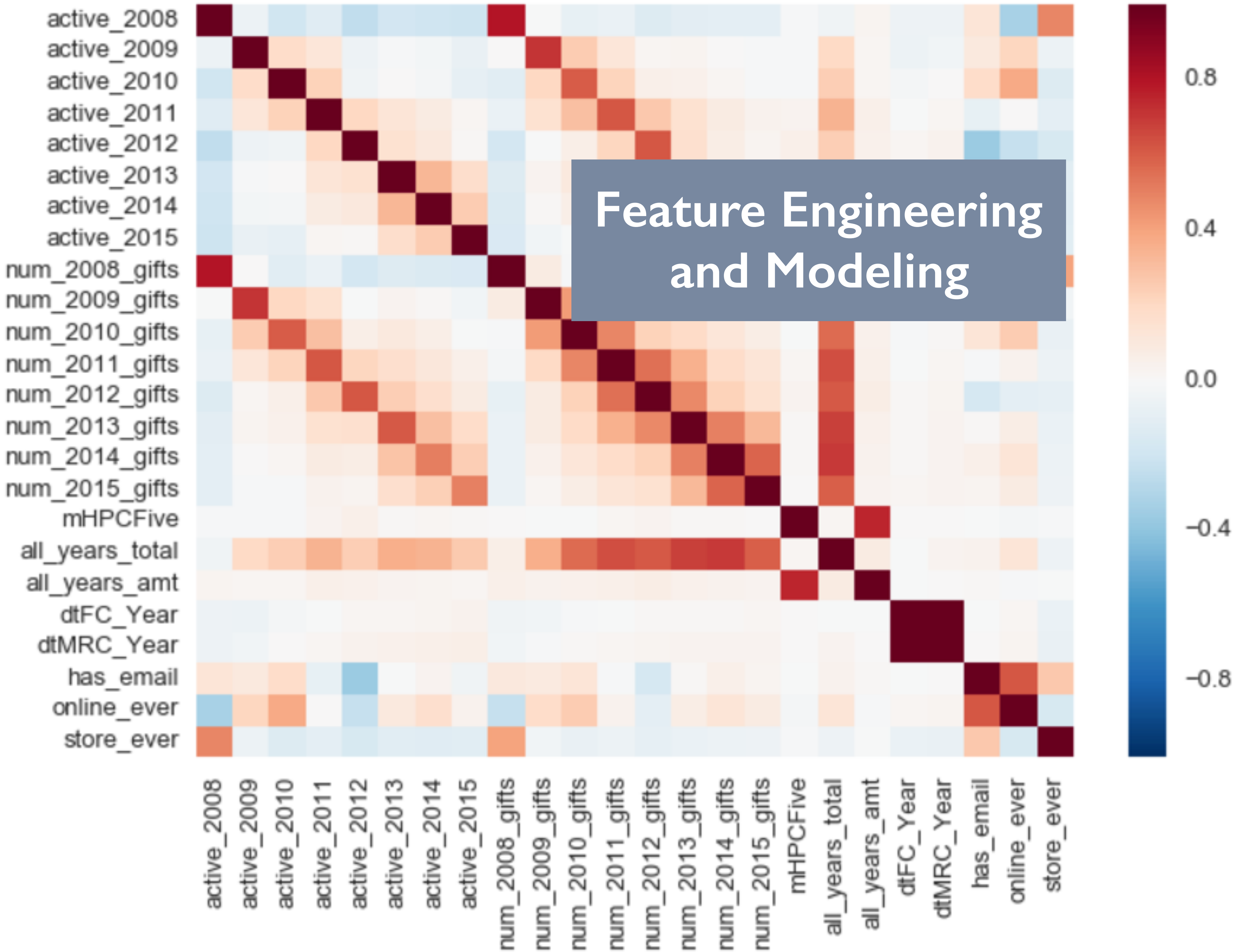
1. Can we predict how long a donor will continue to give to the organization?

2. Are there certain characteristics that make someone more likely to give repeatedly over time?

THE RAW DATA

- All donors whose first gift occurred between 2008 and 2015
- Aggregate Giving History: average gift, gifts per year, gifts per channel, gifts per channel per year
- Indicators for:
    - Online only donors
    - Donors w/ email addresses in the system
    - People who purchased from the Online Store

Feature Engineering and Modeling



FEATURE ENGINEERING

- **Aggregate giving over time, by donor and by channel**
- **Flags for election year only donors, online only donors, mail only donors**
- **Alternatives to Datetimes**
  - Created dummies for month of first gift and year of first gift
  - Calculated months since first gift using timedelta functions, convert to int
- **Zipcode:**
  - Can’t treat as a float or int, really more a categorical
  - Created zip_region indicator off first digit of zip

MODELING

- Data scaled before regression to offset impact of high dollar donors
- Feature set of 11 values, plus first gift month and account dummies

- Linear regression
- Ridge regression
- Baysean Ridge Regression
- Random Forest Regression

findings & lessons learned

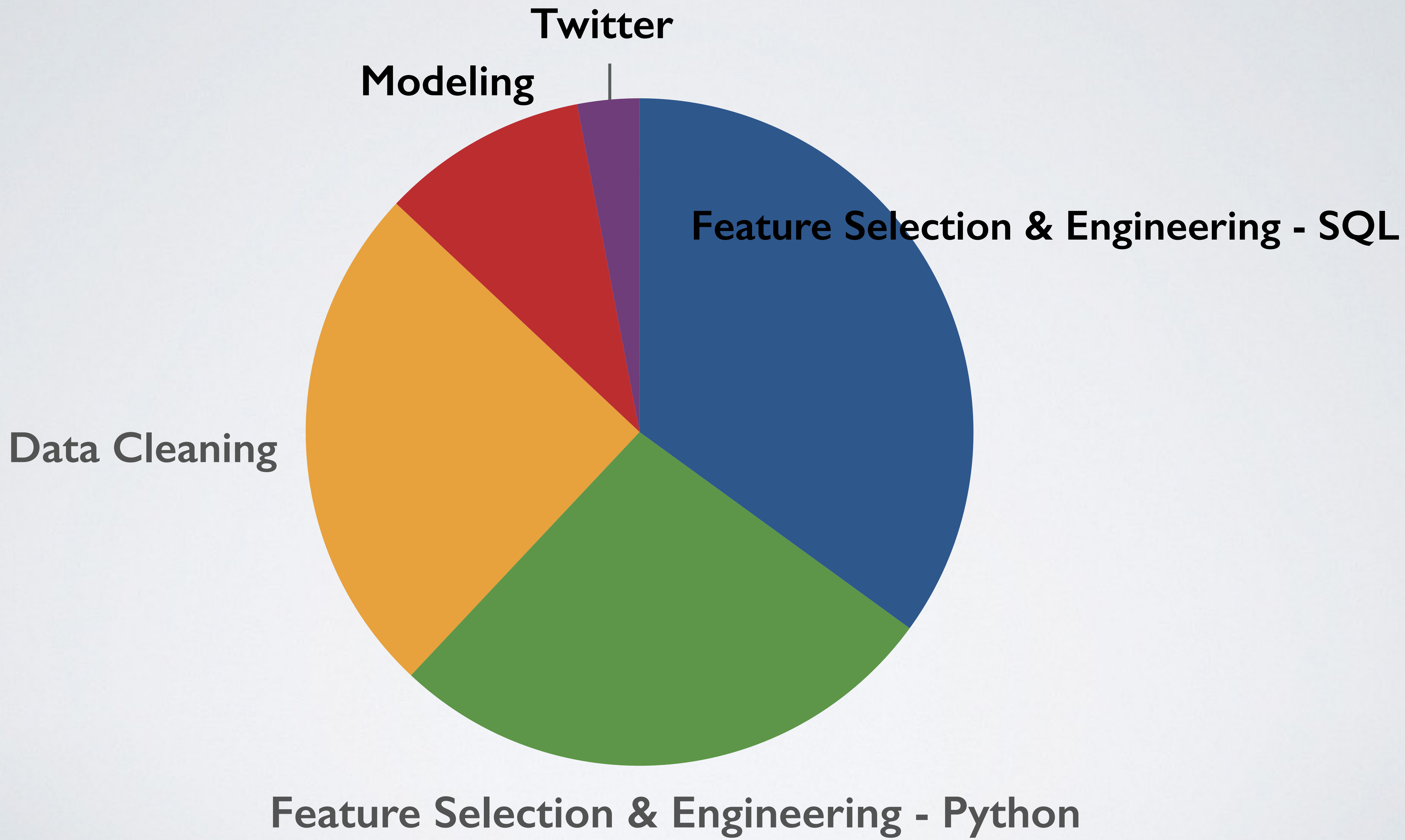
Model	RMSE (10-fold Cross Validation)	Other Params
Null RMSE	1.148532689	
Linear Regression	0.452854407	
Ridge Regression	0.452854406	alpha = 1.0
Bayesian Ridge	0.452841997	
Random Forests	0.522537554	max_depth = 4
Optimized Random Forest	0.235205041	max_depth = 13

Conclusion: Bayesian Ridge performs slightly better than others, but not by a significant amount

```
Skipping line 199174: expected 128 fields, saw 129

Skipping line 258189: expected 128 fields, saw 129

Skipping line 477846: expected 128 fields, saw 129

Skipping line 487703: expected 128 fields, saw 130

Skipping line 488292: expected 128 fields, saw 129
Skipping line 488638: expected 128 fields, saw 129

Skipping line 508883: expected 128 fields, saw 129

Skipping line 570933: expected 128 fields, saw 129

Skipping line 574103: expected 128 fields, saw 129
Skipping line 574984: expected 128 fields, saw 129

Skipping line 606825: expected 128 fields, saw 129

Skipping line 636228: expected 128 fields, saw 129

Skipping line 687399: expected 128 fields, saw 130

Skipping line 794761: expected 128 fields, saw 129

Skipping line 819196: expected 128 fields, saw 129

Skipping line 849668: expected 128 fields, saw 129

Skipping line 856968: expected 128 fields, saw 129

Skipping line 878954: expected 128 fields, saw 129
```

Challenges

HOW I SPENT MY TIME



NEXT STEPS

- Gradient Boosting Regression
- Increased tuning of models
- More feature selection analysis
- Include Telemarketing data
- Include Consumer/demographic data
- Look at coefficients on features

LESSONS LEARNED

Data Cleaning is …

…Frustrating

Beware the Feature Selection Quagmire

Datetime variables are extra tricky

Leave more time than you think you need

Answering the question is only part of the problem