Predicting House Sales

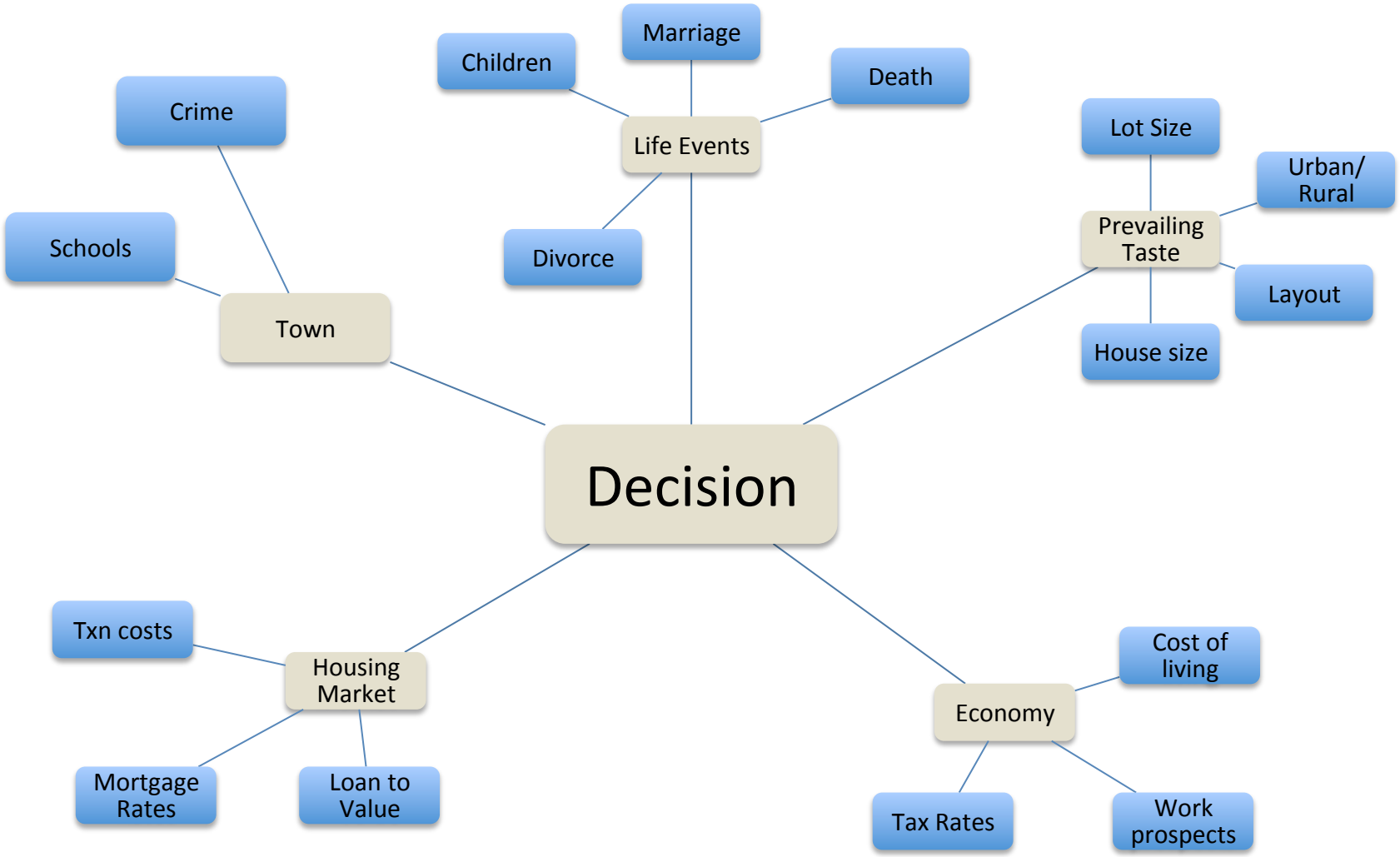
Who will move this year?

Introduction

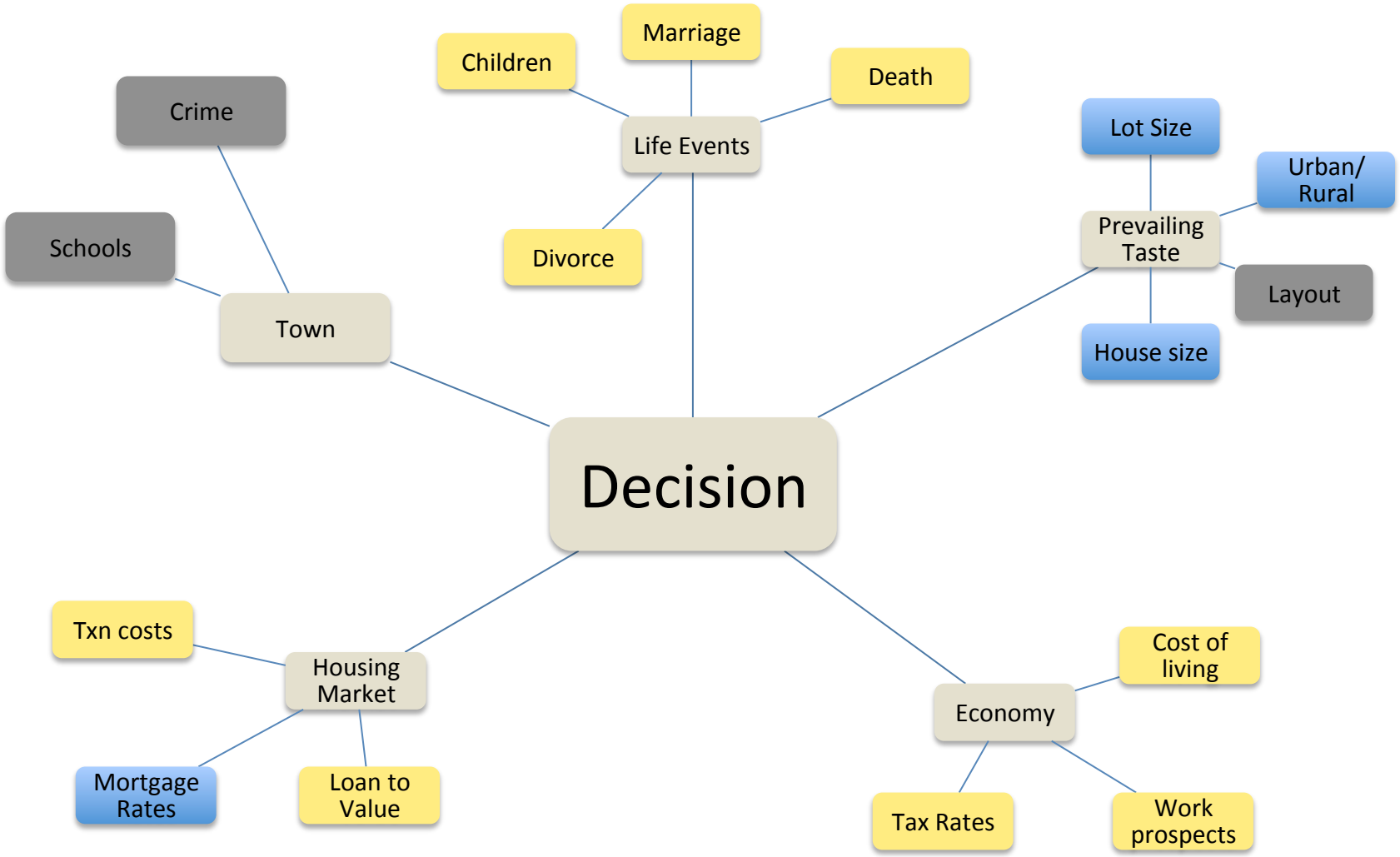
"Half the money I spend on advertising is wasted; the trouble is I don't know which half."
-- John Wanamaker

- SmartZip: spent 7 years building out analytics to answer the question **Who is likely to sell their house in the next 12 months?**
- Project goal: build a competing system in 3 weeks

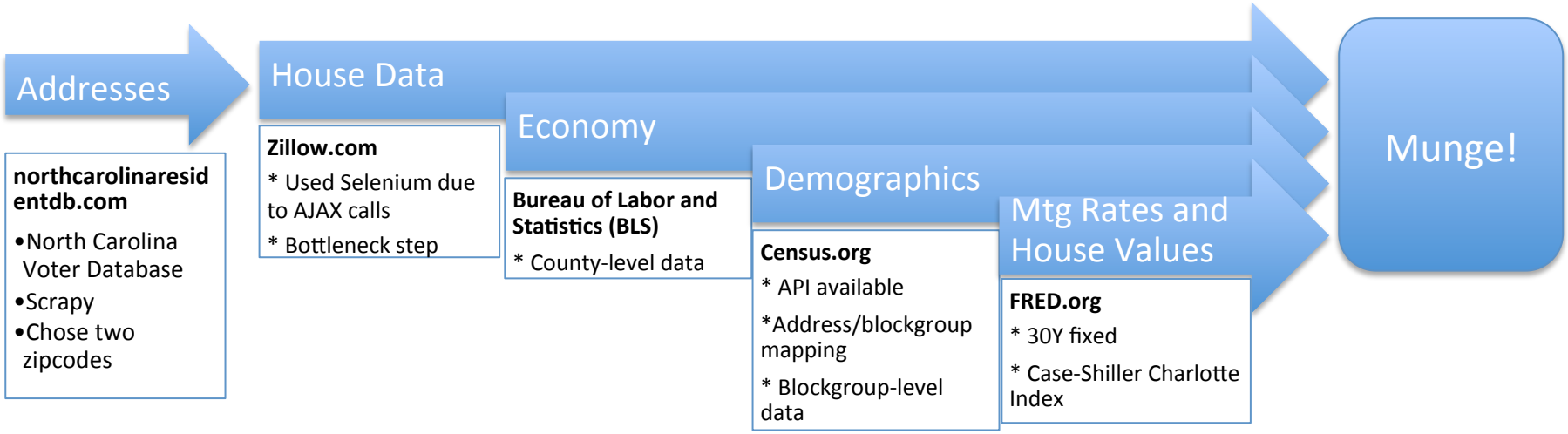
What is considered when deciding to move house?



Had to use many proxies for the model



Data procurement was a hurdle



Bottlenecks:

- Zillow data – Selenium is slow
- Munging – iterative process more labor intensive to build than one-off
- Census data – API easier than data files but has less granular data

Got 10% of the desired observations, but project still viable

- 18,000 observations
  - 2,500 houses
  - 10 years of observations
  - 2 zip codes
- 61 variables => 24 features
- Unbalanced: 5% of houses sell in a given year

Missingness was important

Houses not sold for a long time have less comprehensive data

Missing Item	Type	Action
House profile not accessible on Zillow	MCAR?	Hope and prayer
Square footage	MNAR/ MCAR	Regression on number of bedrooms
Inadequate sale data	MNAR?	Omitted
Address not available in census	MNAR?	Omitted
Others	MCAR	Mean/median

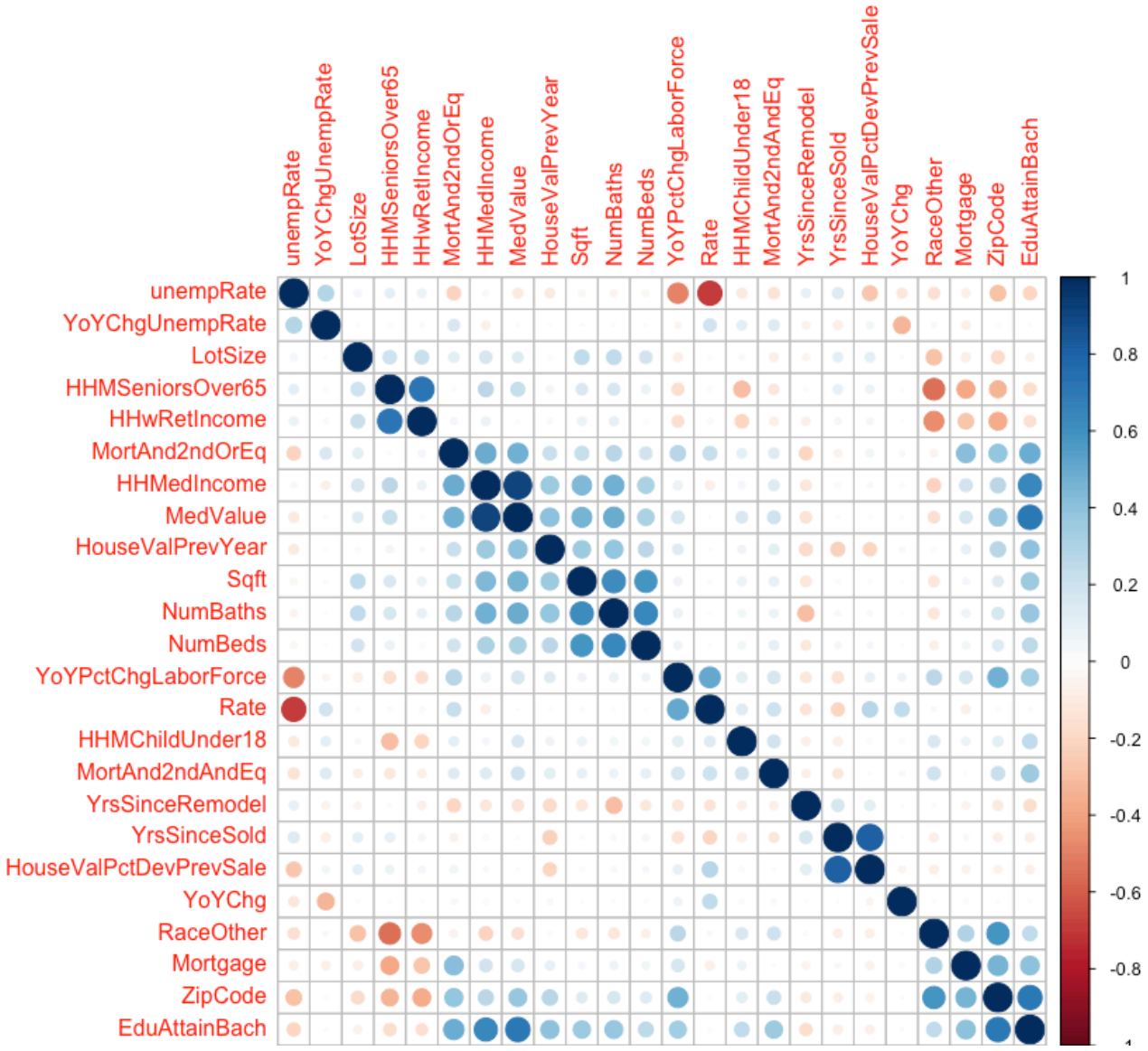
Missingness: house value imputation

- True house values are revealed during a sale
- For non-sale years, had to estimate based on proxies
  - Adjust most recent sale by cumulative log returns of Case Shiller Charlotte Home Price Index
  - Further adjusted based on difference between yearly real home sale values and implied values

Aside: missingness with insidiousness

Sales transaction sparseness increased into the past. Feature engineering inadvertently created bias in data set: dropped observations for which "Years Since Sold" couldn't be derived.

Feature set looks robust…



…but not all features acting intuitively

Group	Feature	More/less likely to sell	Intuitiveness
People	Kids in house	Less	✔
	Non-Caucasian	More	?
	Second mortgage and home equity loan	More	✔
	Moved recently?	Less	✔
	Haven't moved in a while?	Less	✖
Places	Large lot	Less	?
	Large house	Less	?
	House value	More	?
	House worth more than purchase price?	Less	✖
	Zipcode	More/Less	✔
Things	Growth of labor force	More	?
	Unemployment rate	Less	✔
	Increasing mortgage rates	More	✔

We can reduce features, but it may be premature

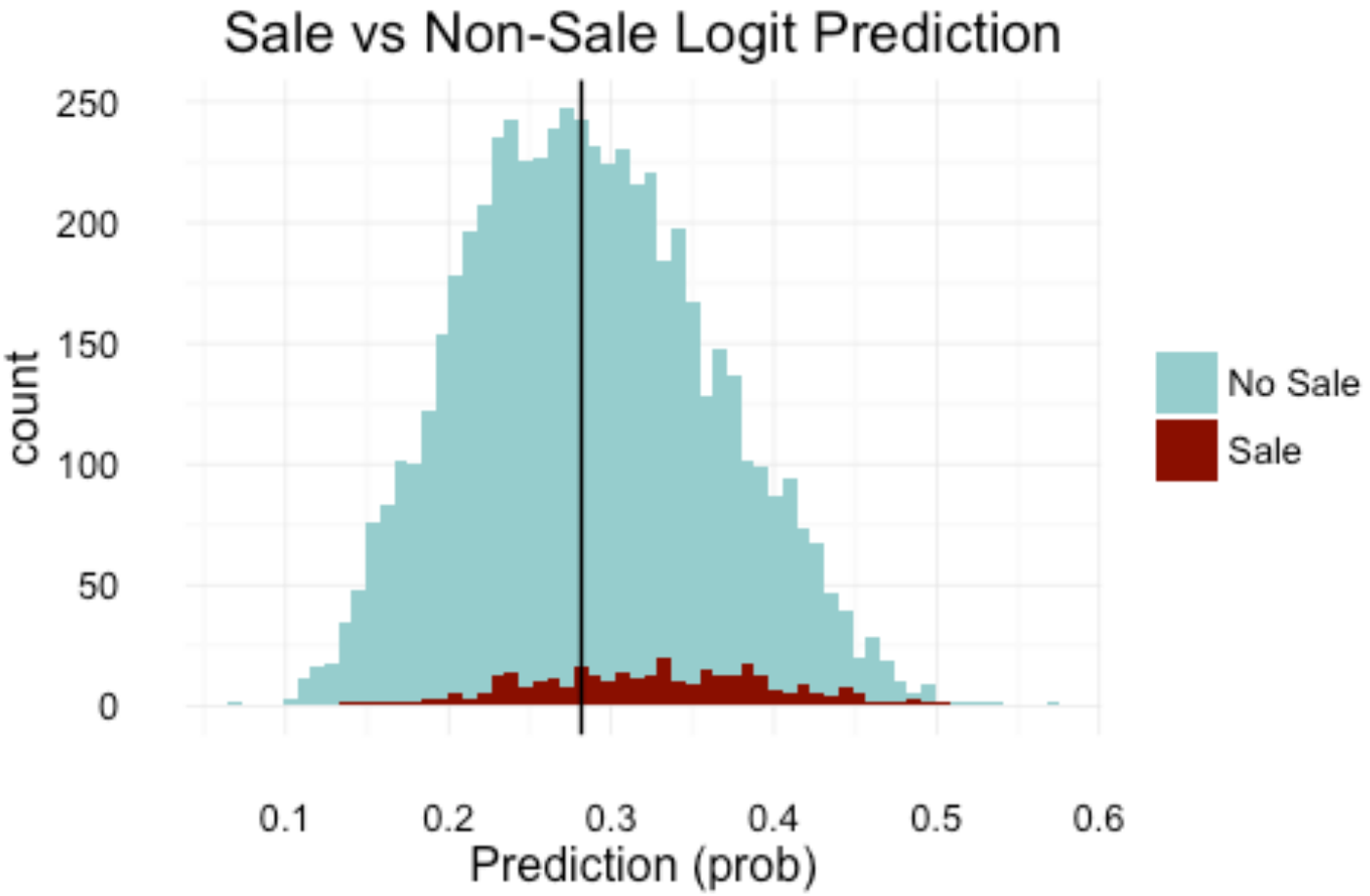
- Models disagree on which features are important
  - Logistic: drop 11 features
  - Lasso: drop 5 features
  - Random Forest/XGB: important features among those logistic and lasso suggest dropping
- Some features may have more utility with more data or more granular data

Summary: it's hard to predict the sale of a house…

Modeling

- **Logistic Regression**
  - Many significant coefficients
  - Little predictive power
- **Random Forest**
  - Classification failed
  - Regression was similar to LR
- **XGBoost**
  - Different distribution, but similar results
- **SVM**
  - Never finished tuning
  - Possibly too much overlap in classes
- **Ensembling**
  - No improvement in ROC

…but the skews in the distributions are helpful



Because we can improve targeting by 60% for half the prospects

Baseline

	Count
No Sale	6128
Sale	328

1 out of 20 homeowners are prospects

Model

	No Sale (predict)	Sale (predict)
No Sale	2902	2898
Sale	98	230

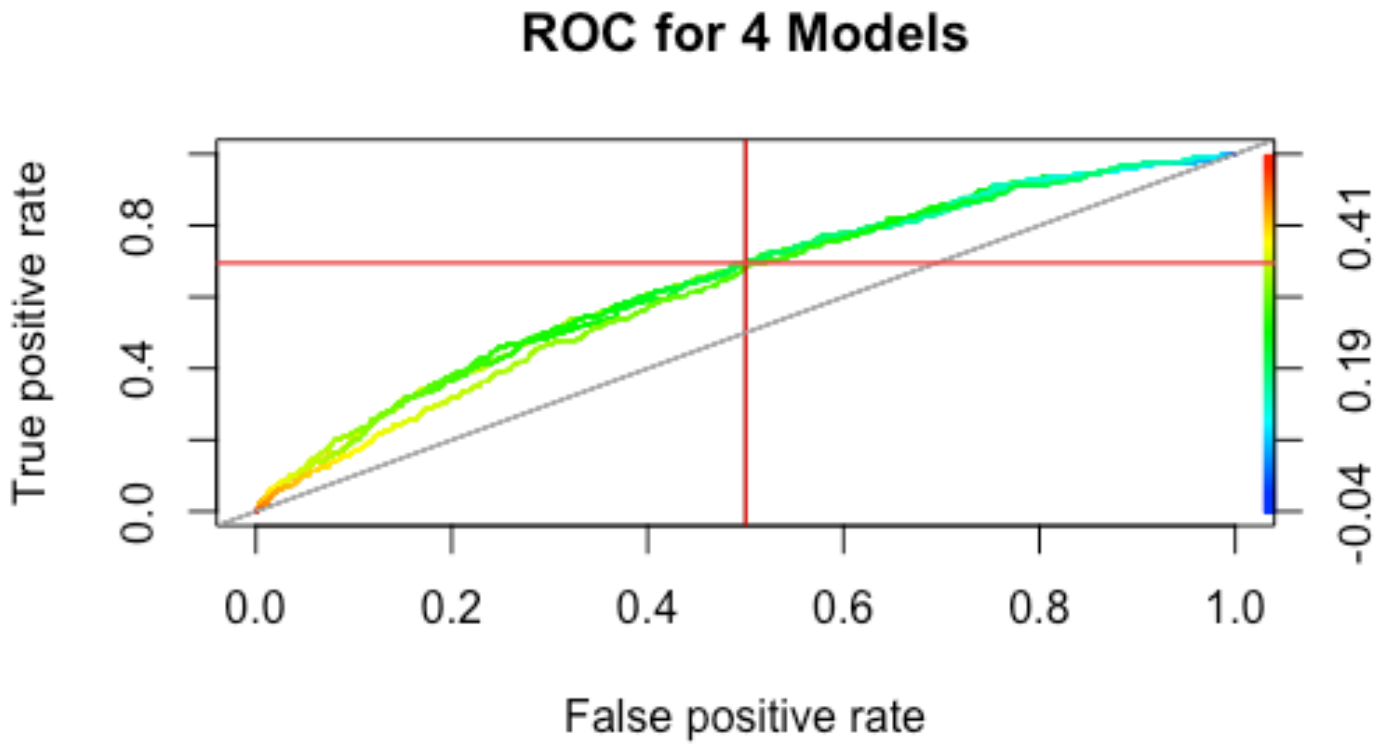
1 out of 13 homeowners are prospects

Next Steps

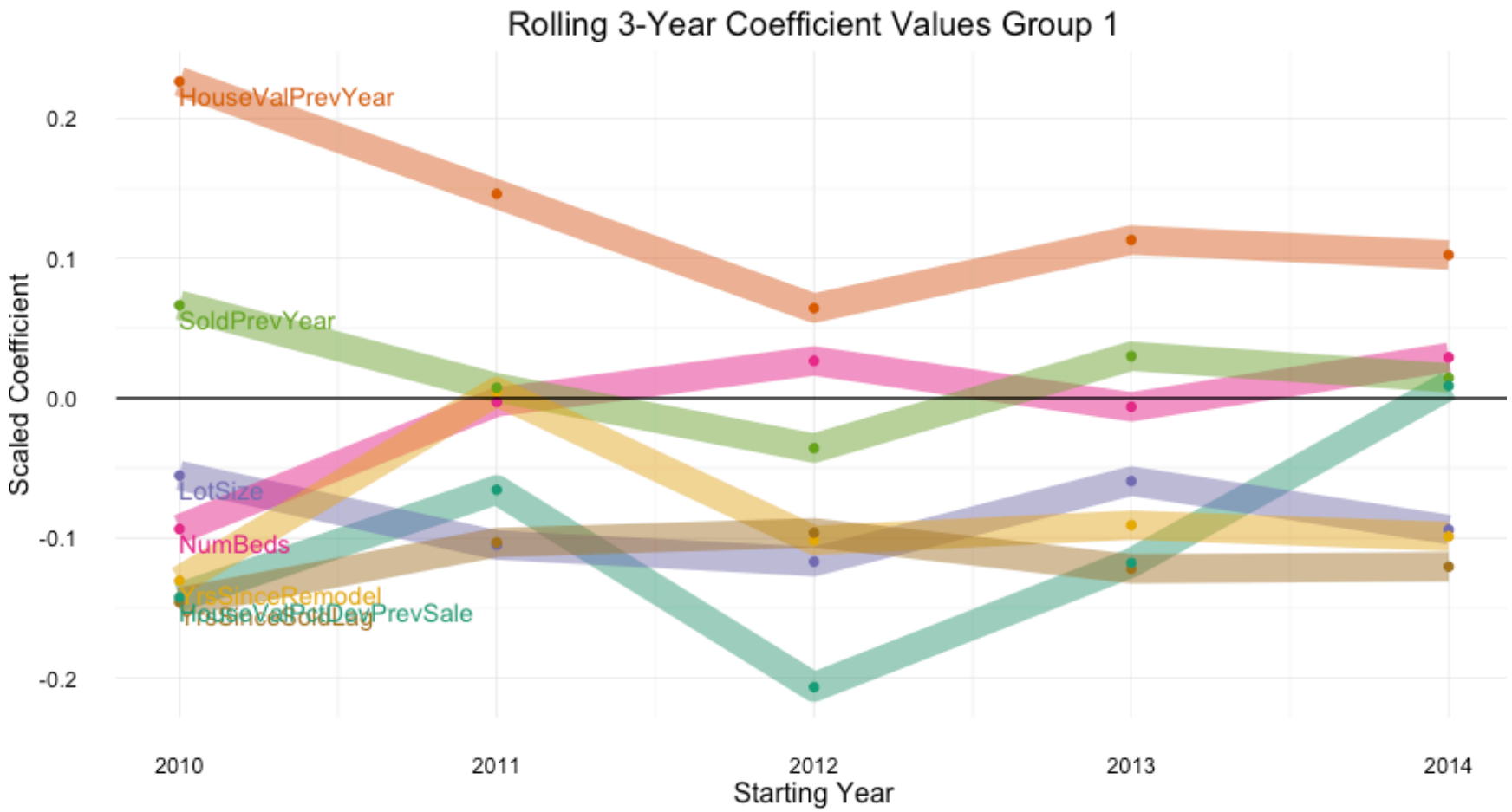
- Go broader rather than deeper
  - Quality of sale data diminishes with distance into past
  - Coefficients not consistent over time
  - Getting more houses will compensate for limitations of census data
- Census block-level data
  - More detailed neighborhood demographic data
  - Better estimates of housing prices
- Mapping feature in Shiny

Questions?

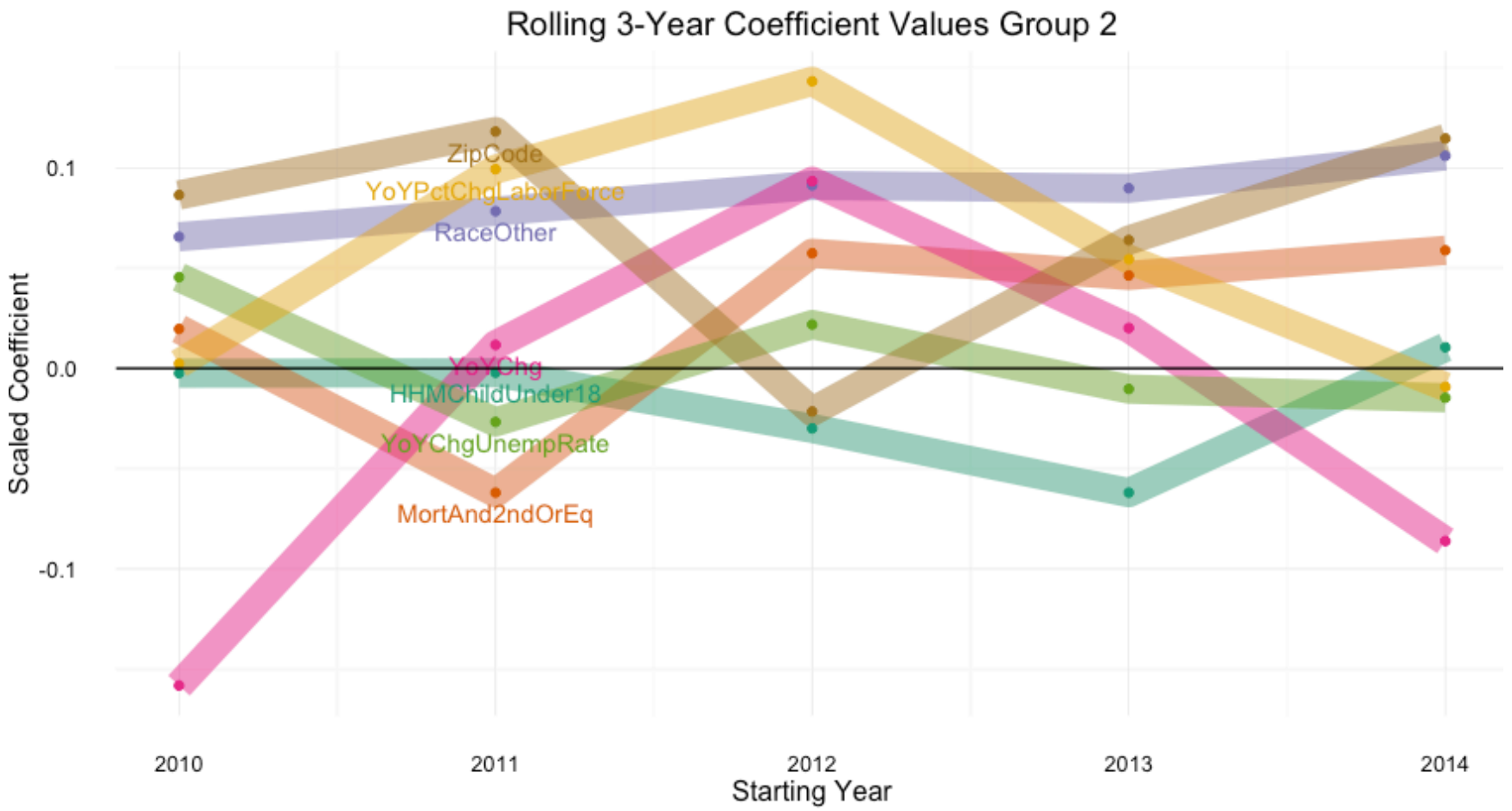
Using 0.5 false positive rate



Coefficient Stability



Coefficient Stability



PC2 is Interesting