NLP in Home Sale Price Prediction

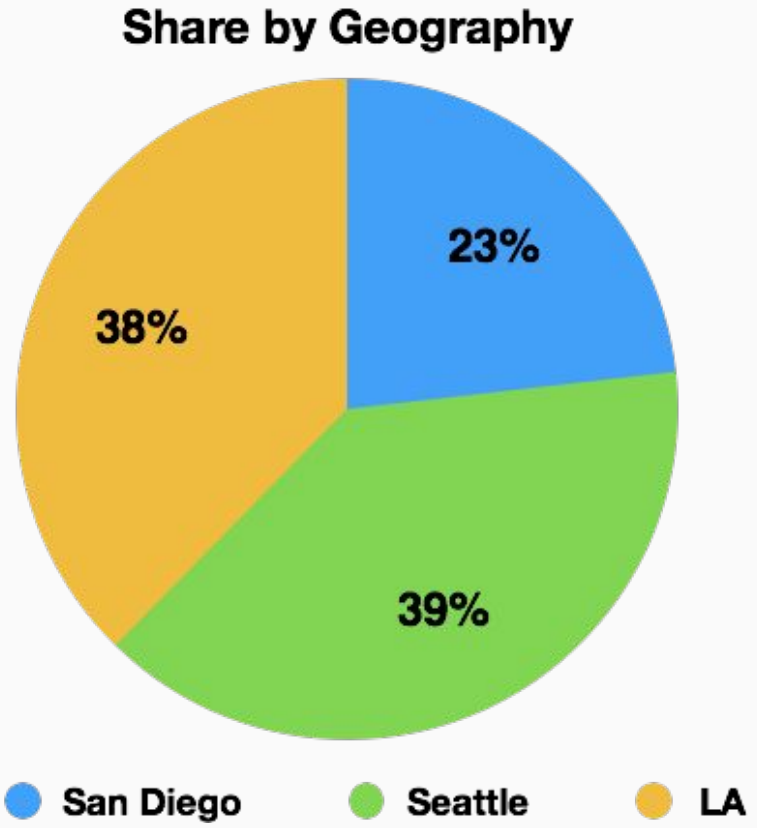
New Data Set

Web Scraped Data

- 25,000 transactions
- 2015 to 2018 time period

Three Market Sample

- Greater Los Angeles
- San Diego
- Seattle



Feature Categories and Examples

Basic House Data	*Beds, Baths, Square Footage, Lot Size, Age of House*
Geographic	*Latitude, Longitude, 3 Market Flags*
Text Features	*10 Standardized Topic Weights*
Zip Code Level Demographics	*Population, Education Income, Public Schools*
Lagged Market Comparables	*$ per Square Foot in Zip Code, Case Shiller National Index*

Sample Text Topic Features from NMF of TF-IDF

View and Location

view deck lake mountain

home outdoor high

expans

Investors and Fixer-Upper

verifi opportun inform sq ft

permit condit investor

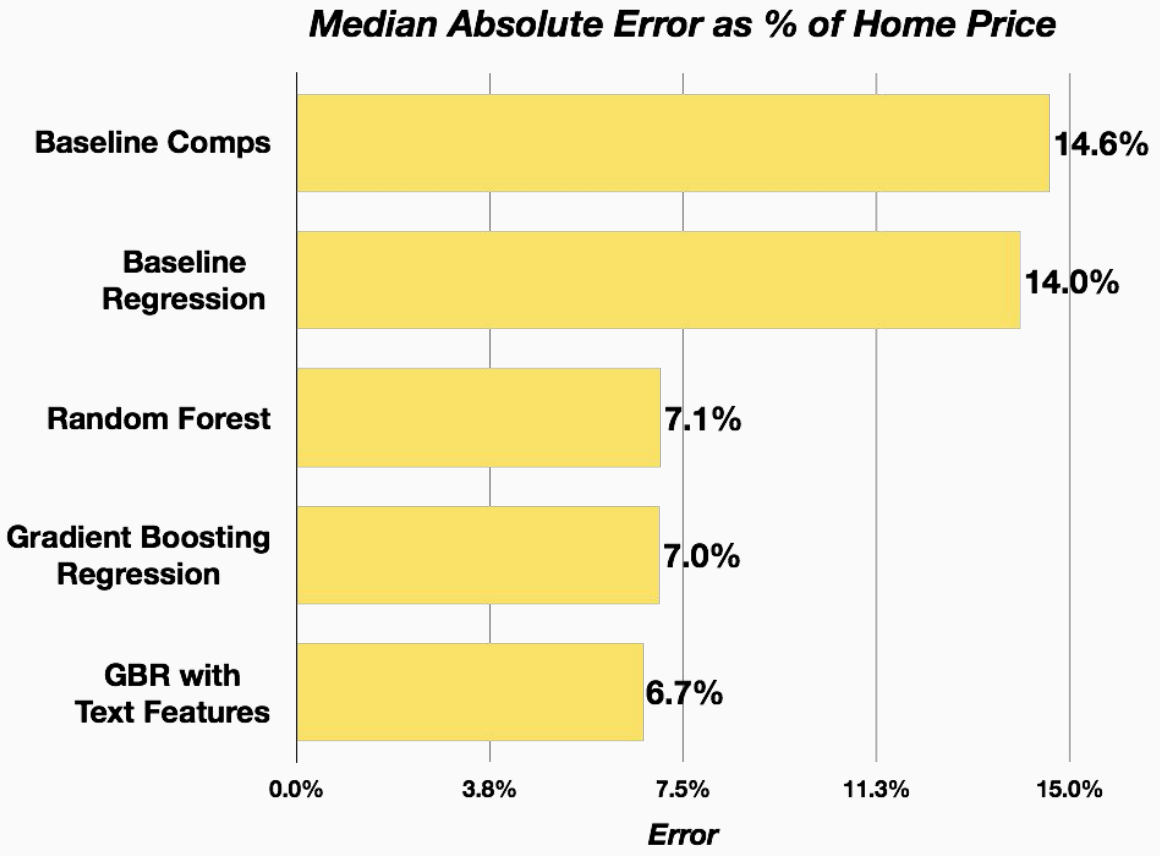
Model Validation and Selection

Evaluation Metric

Median Absolute Error as Percentage of Home Price

Four Models Trained

- Zip Code Comps
- Linear Regression
- Random Forest
- Gradient Boosting Regression



Gradient Boosting Regression on 20% Holdout Set

Holdout Method

Most Recent Transactions

Median Error

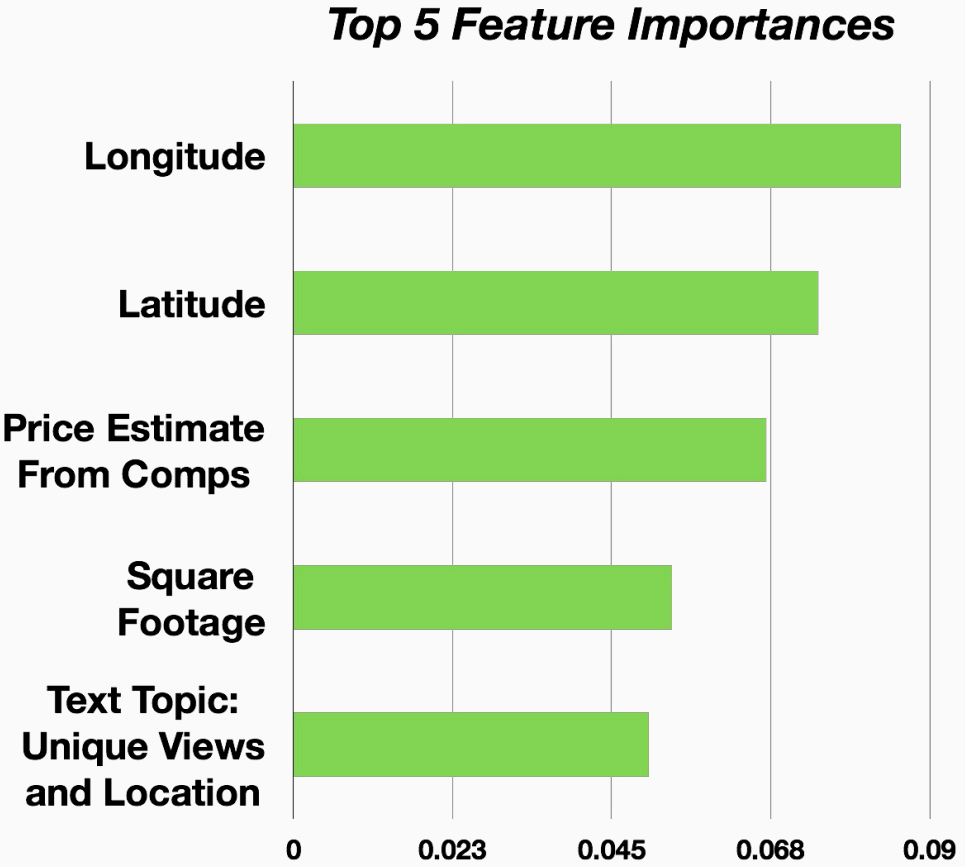
7.9% of Home Price

Text Topics in Top 10

Unique Views / Location

Investors / Fixer-Upper

Yard/Garden



Conclusions and Extensions

Some evidence of text handling extremes in quality

Pursue additional topic models and text features

Thank You

CONTACT INFORMATION

whetfield

@

everything