Insights of the Omaha-Lincoln-Des Moines Area House Prices

Jun Dai

Introduction

- House buyer:

  Which venues to look at?

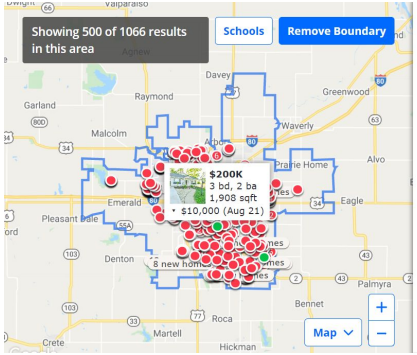
- House seller:

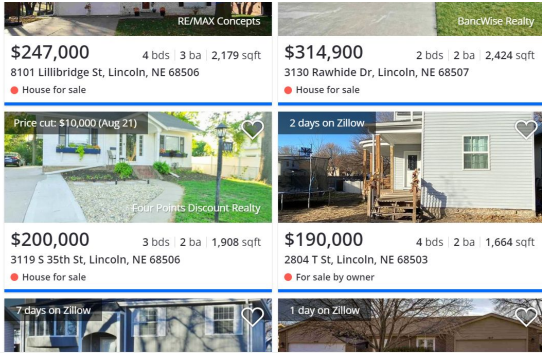
  Which venues (business) to invest to increase the house price?.

- Zillow: One of the largest, most trusted marketplaces.

Data & Features

Final dataset Dimension:
3044*485



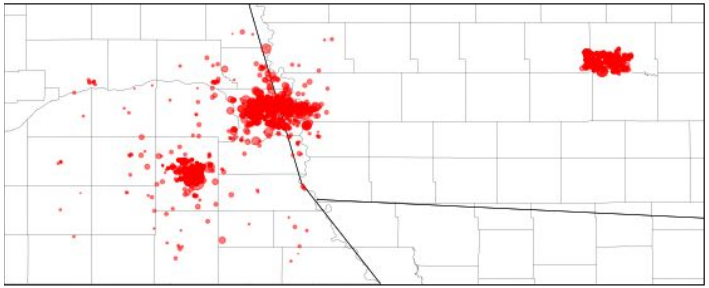


bs4+request
Address, price, link

1. **bs4+request**
   Num_bed, num_bath, flooring, roof, size, garage, year_built…
2. **Foursquare, nearby venues**

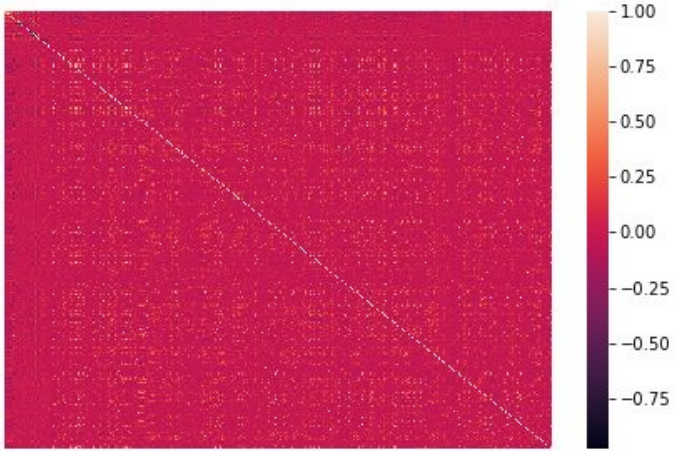
Cleaning & engineering:

Outlier removal,
Unit unifying,
Text to numerical,
Data merge,
...



Processed data

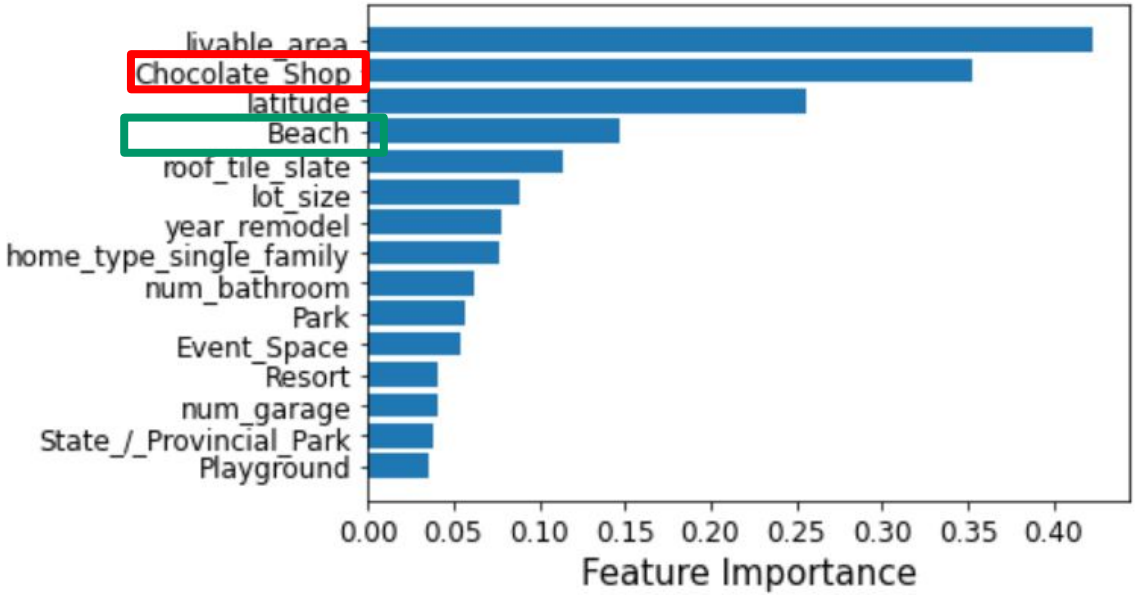
ude	longitude	...	Laundromat	Intersection	Tennis_Court	Locksmith	Monument_/_Landmark	Soccer_Field	Bike_Shop	Hardware_Store	Plaza	prices
022	-96.223930	...	0	0	0	0	0	0	0	1	0	529500.0
270	-93.688860	...	0	0	0	0	0	0	0	0	0	132900.0
598	-93.627449	...	0	0	0	0	0	0	0	0	1	358500.0
910	-93.819442	...	0	0	0	0	0	0	0	1	0	271900.0
938	-96.013697	...	0	1	1	0	0	0	0	1	0	165000.0



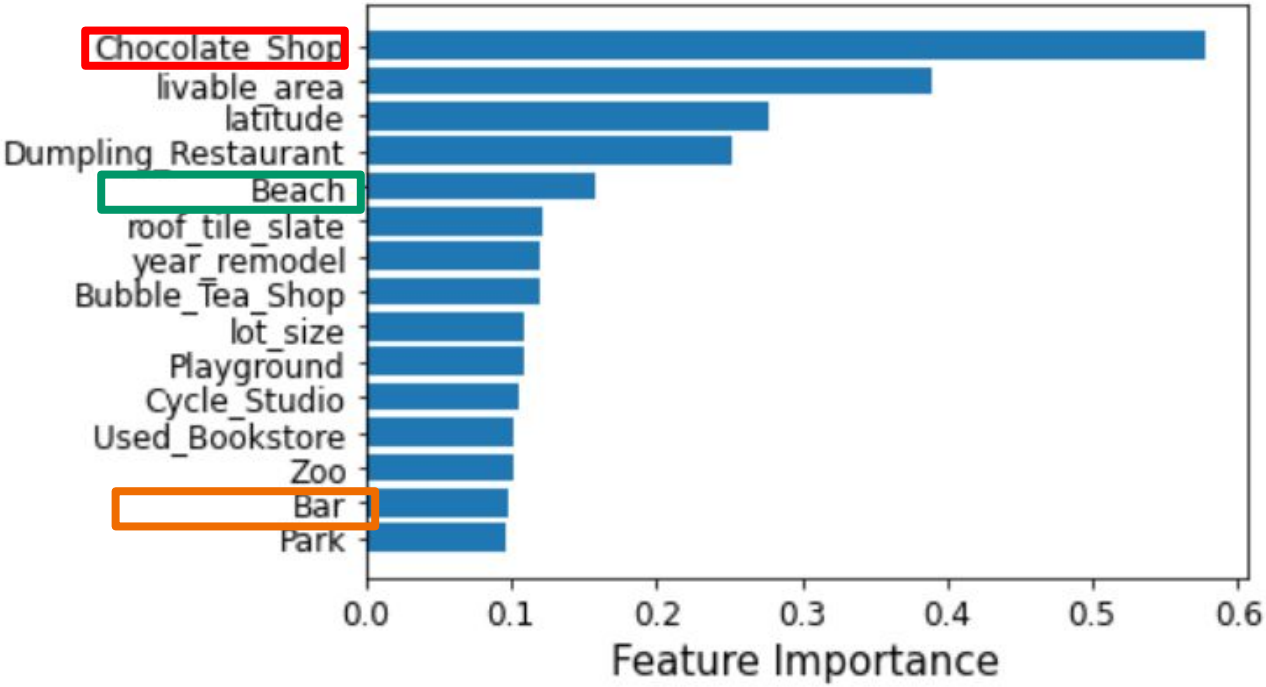
- **Sparse Data**
- **Multicollinearity**

Models:

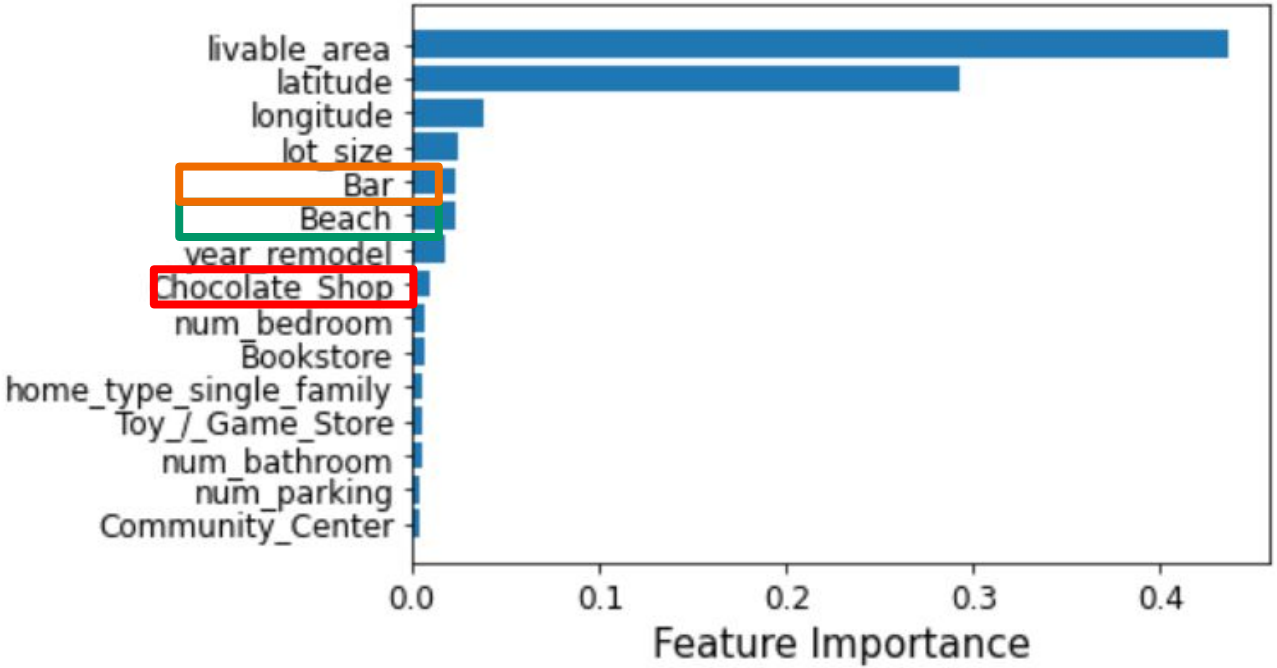
- **Lasso + GridSearchCV**



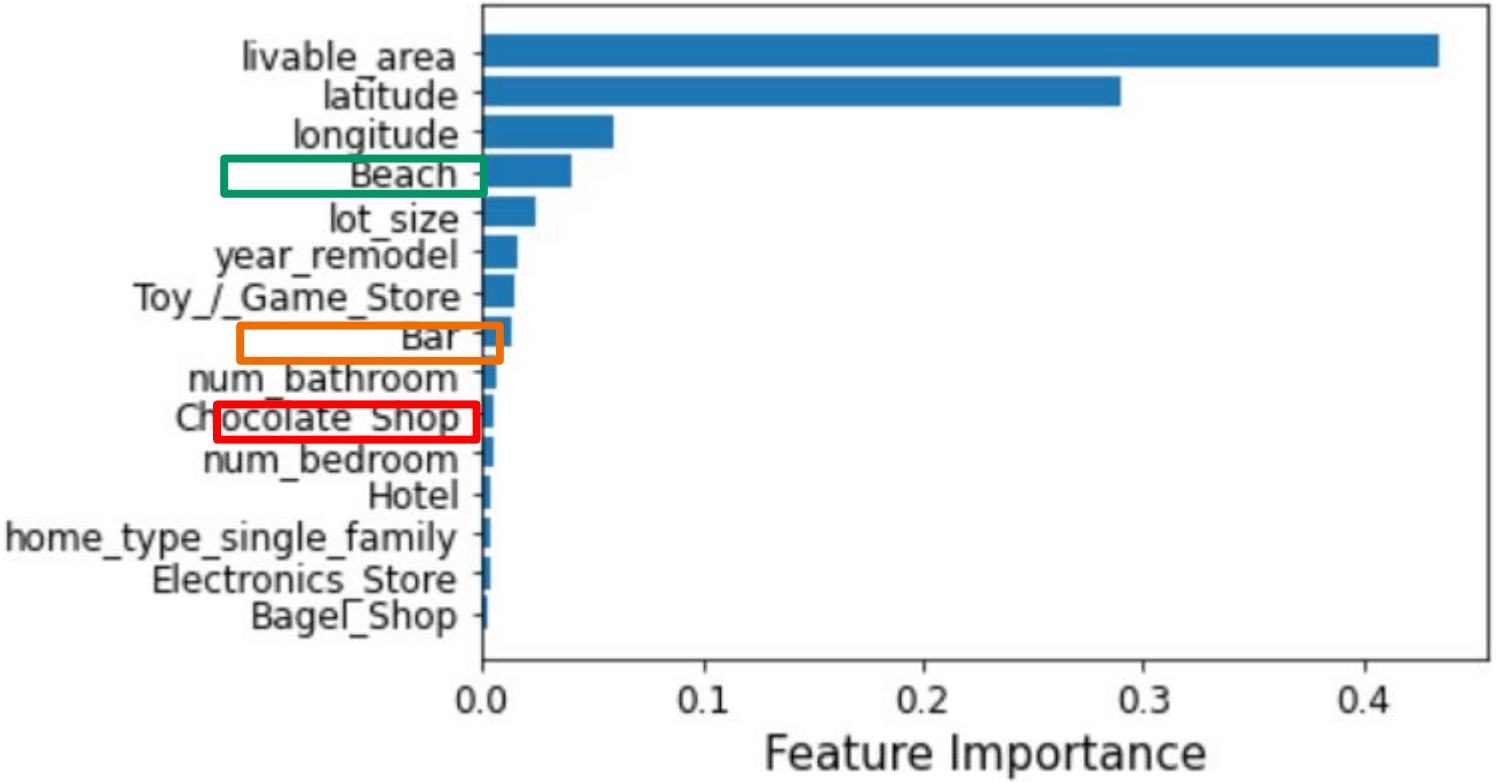
Ridge+GridsearchCV



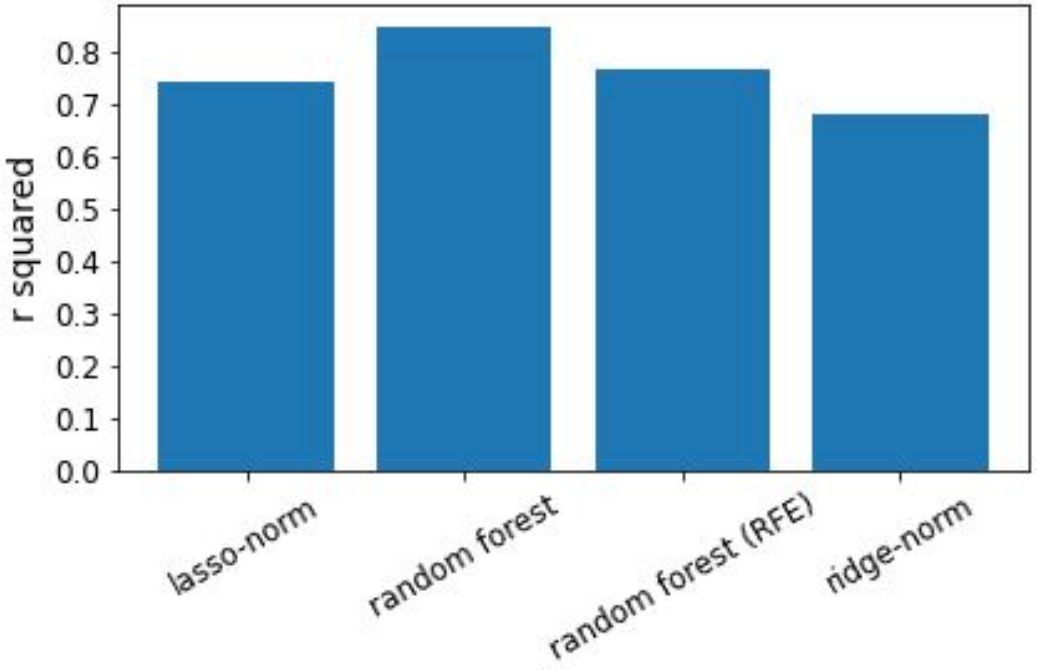
Random Forest + RandomSearchCV



Random Forest + recursive feature elimination



Comparison of four models



Summary

- I scraped a zillow data set for the Omaha-Lincoln-Des Moines area.
- I complemented the zillow data with Foursquare nearby venues.
- Used Lasso and Random forest for regression. Random forest holds the best performance.
- Although there is inconsistency between the feature importance from different models, “Chocolate shop” and “Bar” appear to be interesting features show up in all models.

- **Future work**:
  - A closer look is needed to check the relevant features such as Chocolate shop, yoga studio to confirm the finding through statistical analysis.
  - Model parameters needs to be further optimized.
  - Try XGboost and other feature selection methods.

Thank you!