Predicting real estate prices

Data, methods, value

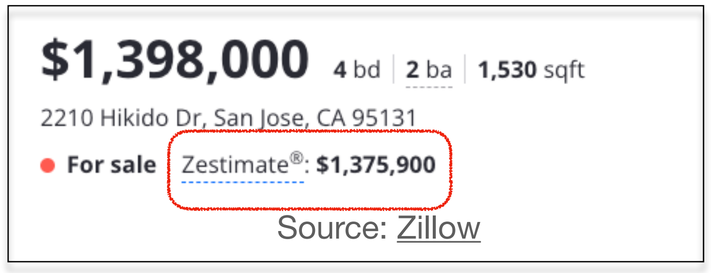
1. What data do we need to predict property prices?

2. And how accurate our predictions will be?

Price Estimator

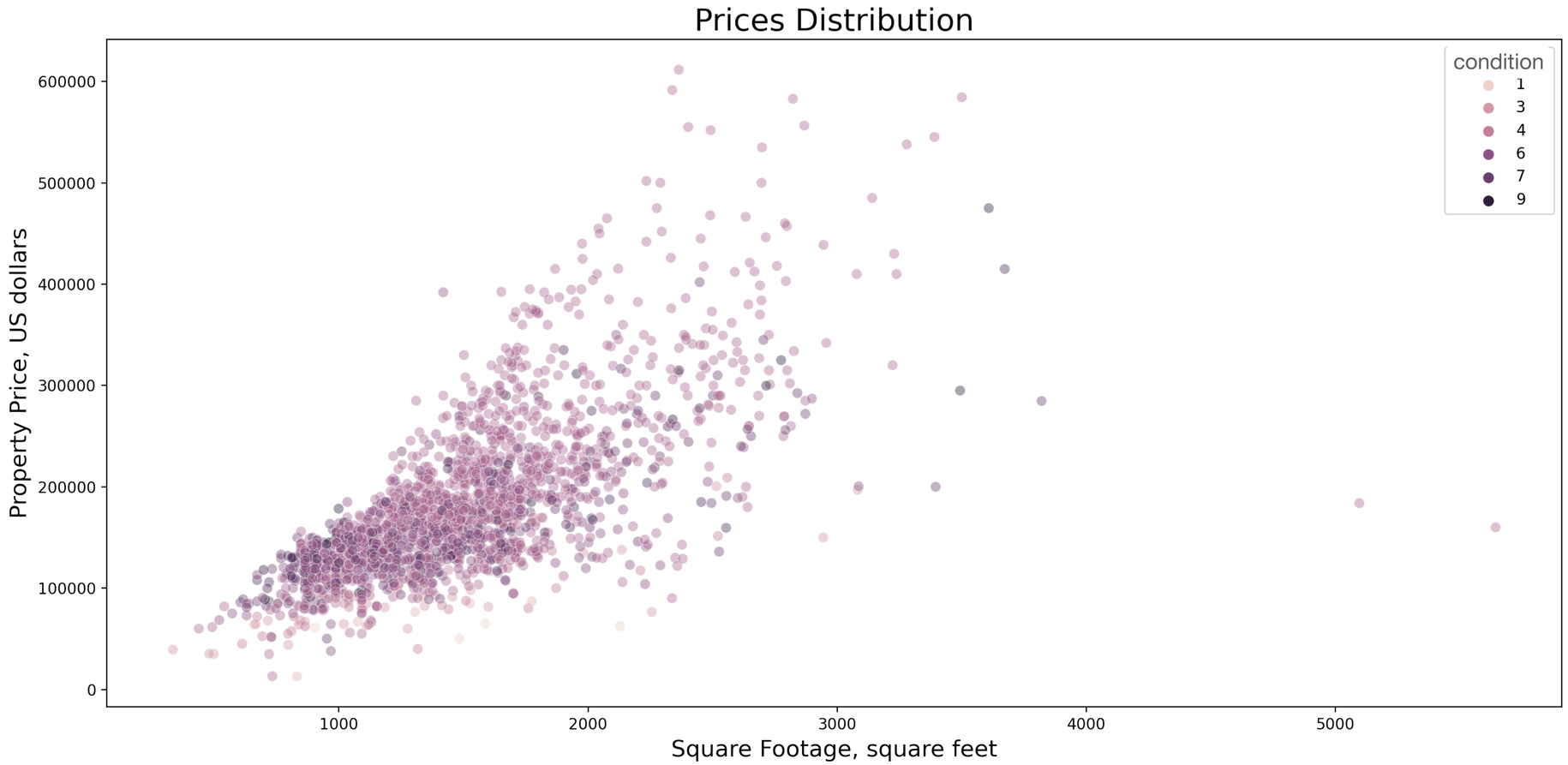
Built into a real estate listing app / website

- Helping buyers and their agents to set the property sell price
- Making a property more or less attractive for buyers based on the real price <> estimated price difference



Data

- 2051 rows x 81 columns
- Ames, Iowa
- 2006 - 2010



Data filtering

Factors to consider

- Is this feature important when predicting property price?
- How values are distributed?
- If the data is categorical - does price differ between classes? To what extent?
- Are classes representative enough to be dealt with?
- Is this feature related to any of other ones?

Data transformations

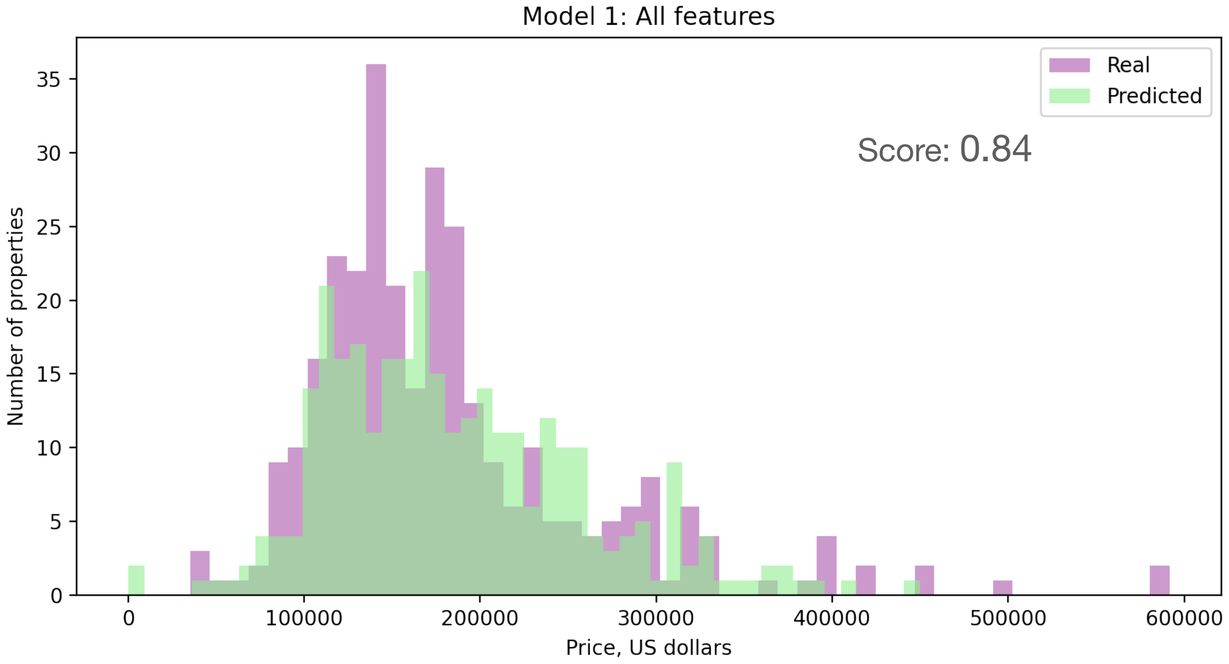
- Remove
- If data brings value into price prediction:
  - Leave as it is
  - Merge with another attribute
  - Transform
    - Descriptive to score
    - Categories rearrangement
    - Categories to binary

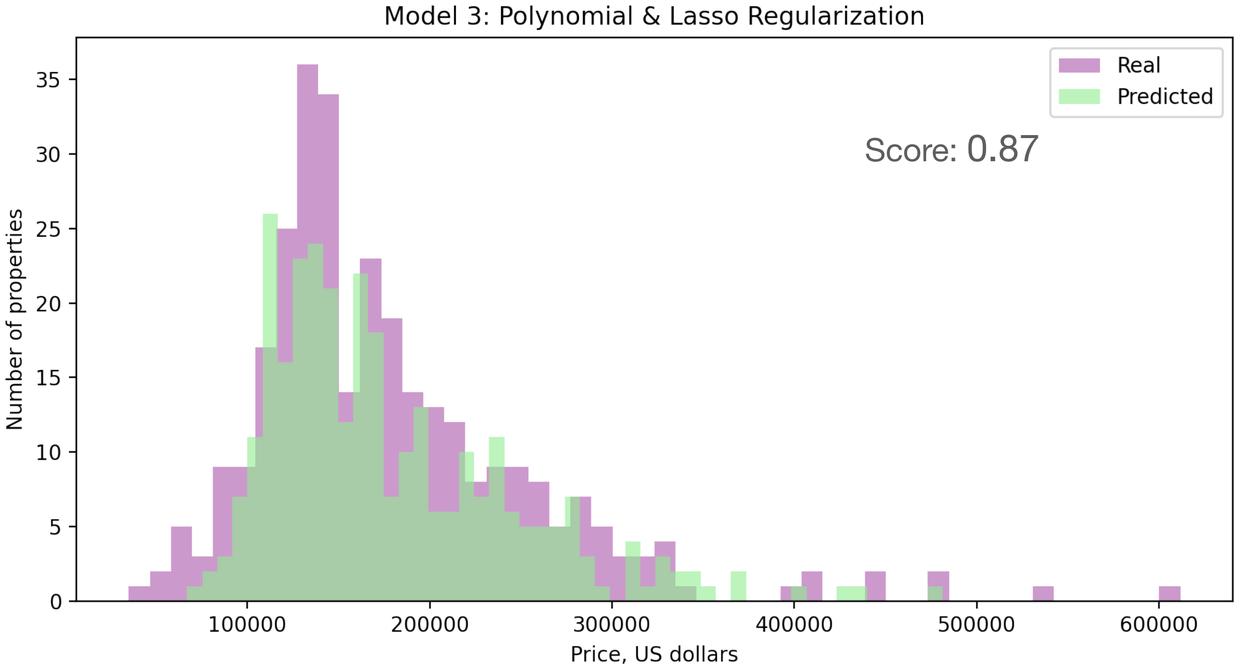
Shortlist

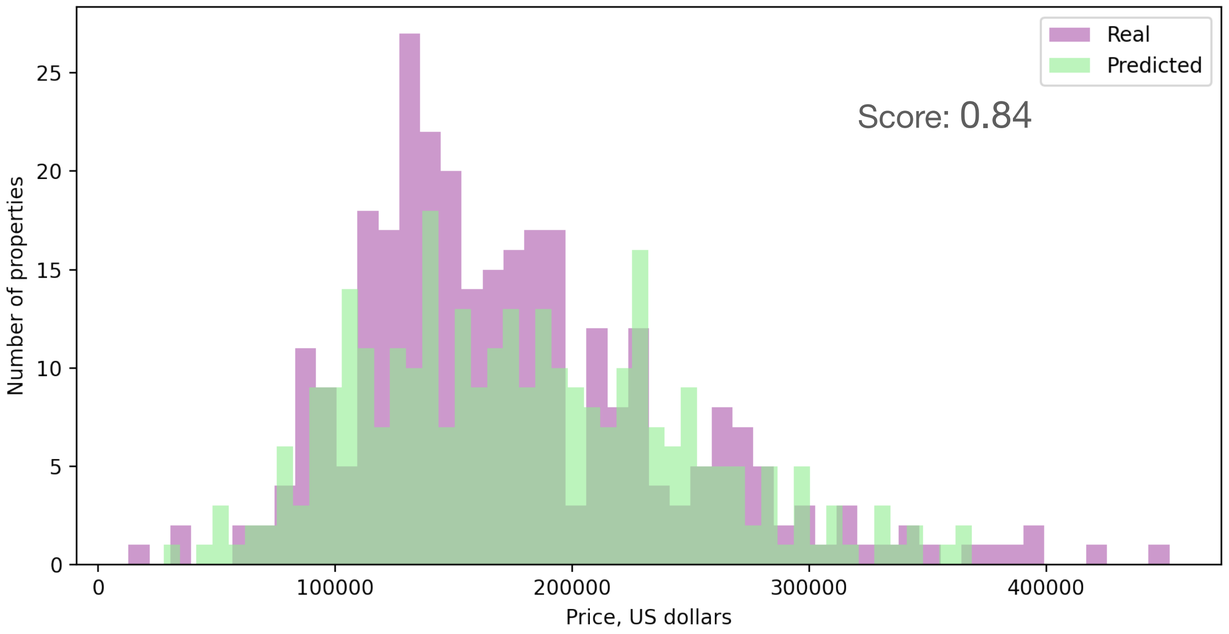
81 -> 18

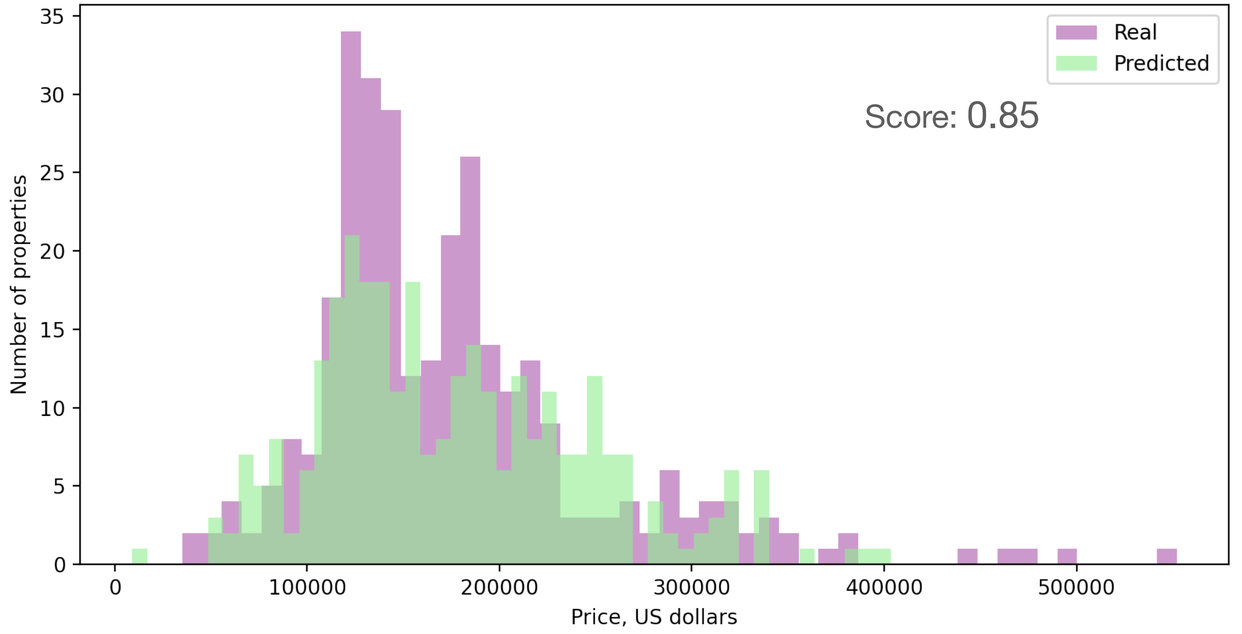
- Id
- Lot area
- Land slope
- Central AC
- Living Area
- Rooms
- New property
- Paved
- Irregular lot
- Cul de Sac
- Positive Feature
- Noisy
- Years old
- Non-living area
- Baseline
- Condition
- Floating Village

Relationships


Model 1: All features
Real
Predicted
Score: 0.84
Number of properties
Price, US dollars
0
5
10
15
20
25
30
35
0
100000
200000
300000
400000
500000
600000


Model 3: Polynomial & Lasso Regularization
Real
Predicted
Score: 0.87
Number of properties
Price, US dollars
0
5
10
15
20
25
30
35
100000
200000
300000
400000
500000
600000


Model 2: Non-collinear features
Real
Predicted
Score: 0.84
Number of properties
Price, US dollars
0
5
10
15
20
25
0
100000
200000
300000
400000


Model 4: Noncollinear Features + Ridge Regularization
Real
Predicted
Score: 0.85
Number of properties
Price, US dollars
0
5
10
15
20
25
30
35
0
100000
200000
300000
400000
500000

Conclusions

Key factors:

- Square footage
- Neighborhood
- Condition

How precise are predictions:

- ~ 33 000 US dollars offset
- 87% of variability of price can be predicted based on this input