Zillow: Sale-Price Model

Mark Dowicz

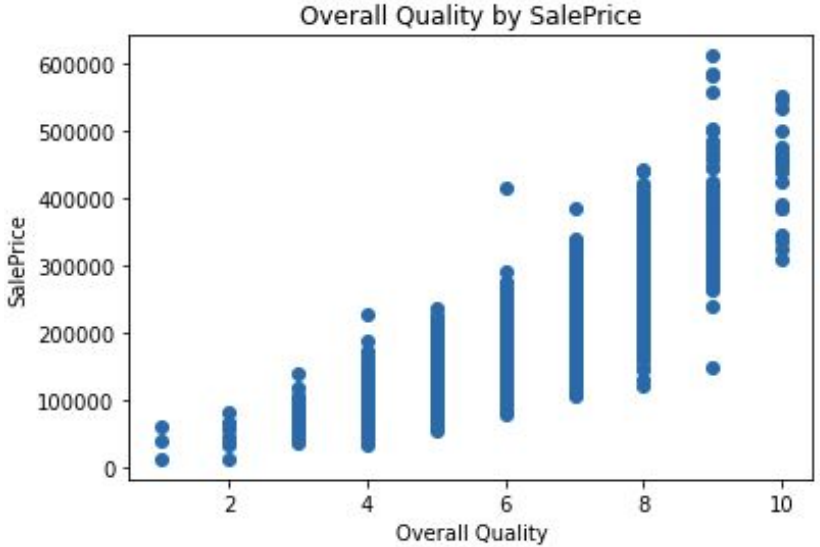
Problem Statement

As a Data-Scientist for Zillow, my team is working on the production of a built-in feature meant to recommend a “fair” sale price for potential home-sellers. My role is to construct a linear-model that will run the back end operations of the feature, which will receive inputs manually entered by the prospective home-seller, to produce said recommendation.

Data Cleaning & EDA

- Numeric Nulls
  - Changed to ‘0’
- Object Oriented Nulls
  - Changed to ‘None’
- Investigated each feature individually
  - Created list of features with visible correlation to sale price
- Among selected features
  - Checked for collinearity, normality
- Removed Visible Outliers
  - Lot Area > 30,000 : Total Bsmt SF > 3,000 : Full Bath >=4 : Total Rooms Abv Ground >= 13

What I'm Looking For



Base Features

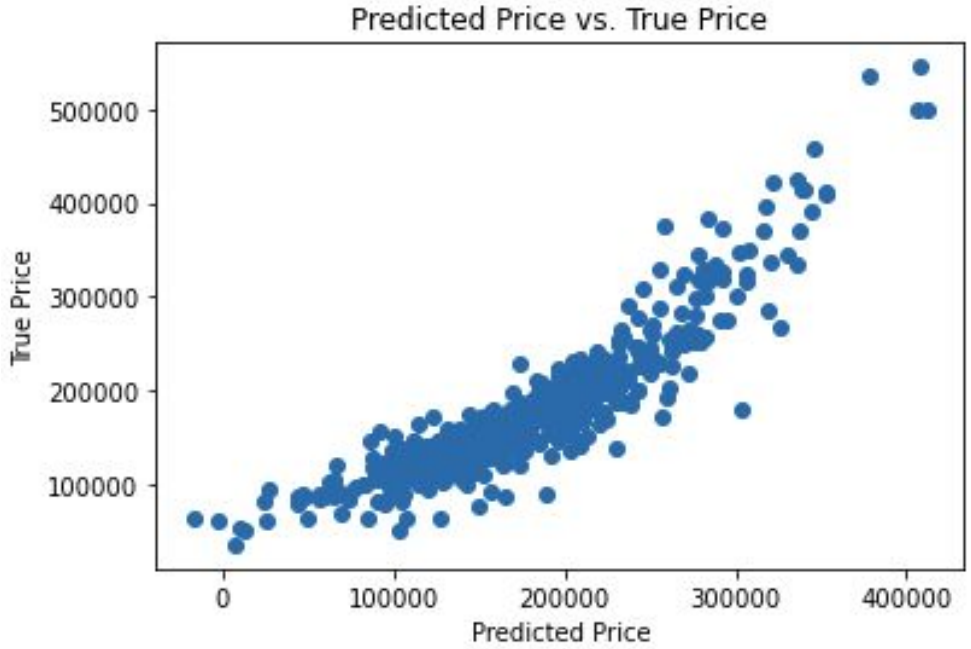
- Quality
  - Overall Quality, Above Grade Living Area
- Space
  - Basement SF, Total SF (1st+2nd Floor), Garage SF, Rooms Above Ground, Lot Area
- Condition
  - Exterior Condition
- Amenities
  - Full Bath, Fireplaces, Fence, Building Type

Modeling: Round 1

	Coefficient	Feature
0	21483.655734	overall_qual
1	39.935946	total_bsmt_sf
2	-7.204637	gr_liv_area
3	3916.464830	full_bath
4	5644.374510	fireplaces
5	52.510564	garage_area
6	-838.289125	totrms_abvgrd
7	1.730060	lot_area
8	53.764301	total_sf

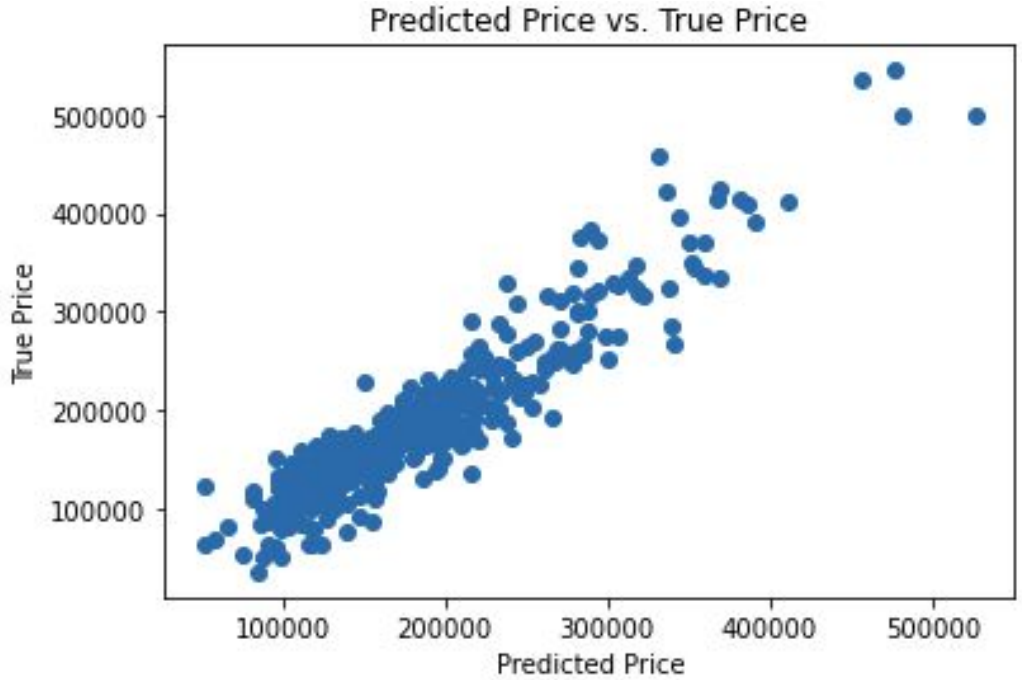
	Coefficient	Feature
10	-18175.611022	bldg_type_Duplex
11	-3356.512351	bldg_type_Twnhs
12	1653.794969	bldg_type_TwnhsE
13	-6307.668694	exter_cond_Fa
14	236.465121	exter_cond_Gd
15	13580.117763	exter_cond_Po
16	-2029.402013	exter_cond_TA
17	2201.663259	fence_GdWo
18	-2881.559678	fence_MnPrv
19	-2496.608010	fence_MnWw
20	826.681619	fence_None

- Basic LR --->
- Train Score: 0.8226
- Test Score: 0.8322
- RSME: 31,937.38


Predicted Price vs. True Price
True Price
500000
400000
300000
200000
100000
0
100000
200000
300000
400000
Predicted Price

Modeling: Round 2

- Created Pipeline
  - Polynomial Features
  - Standard Scaler
  - Lasso
    - Tested Alpha's between 0-10
- Grid Search --> Best Estimator
- Best Estimator: Polynomial Features, Standard Scaler, Lasso (Alpha=10)
- Train Score Test Score RMSE
  - 0.8977 0.8814 26121.17


Predicted Price vs. True Price
True Price
Predicted Price
500000
400000
300000
200000
100000
100000
200000
300000
400000
500000

Conclusions

- Blackbox Model
  - Coefficients are uninterpretable
- Relatively balanced Bias-Variance Tradeoff
  - Train and Test Scores close together
  - Only 1 feature per house feature
- Predicted Values:
  - 2 of the 878 unseen values were negative!
- Implementation?
  - Not yet

Issues

- Year and Economic environment (interest rates) not accounted for
- House Age and Renovations not accounted for
- Unidentified coefficients too strong
    - Blackbox Model
- Implementation
    - User Error
    - Scale

Thank You!