Linear Models of Housing Prices in Ames, IA

•••

Interpretability and Predictive Power

Background & Data Science Problem

- In 2021, the total value of US real estate grew by nearly 19%, to a value of $43.4 trillion[1]
- Furthermore, a plurality of US citizens believe that investment in real estate is the best way to build personal wealth[2]

❖ A group of real estate investors wants us to build a model to predict sale prices of homes in Ames, IA.
  - They care primarily about the predictions being *not too far from correct*
  - They also want to be able to *understand* which features of homes contribute to higher predicted sale prices

---

1. https://www.zillow.com/research/us-housing-market-total-value-2021-30615/
2. https://www.cnbc.com/2022/12/15/americans-say-real-estate-is-best-way-to-build-wealth.html

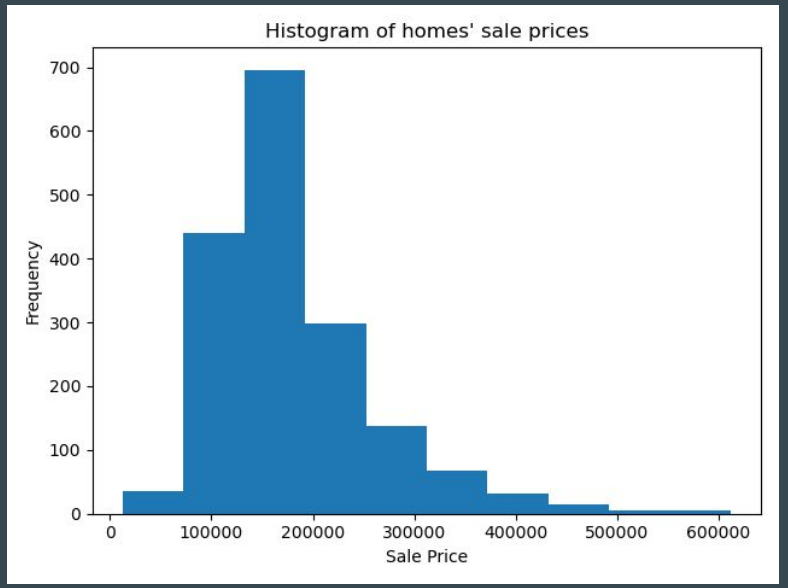
Methods

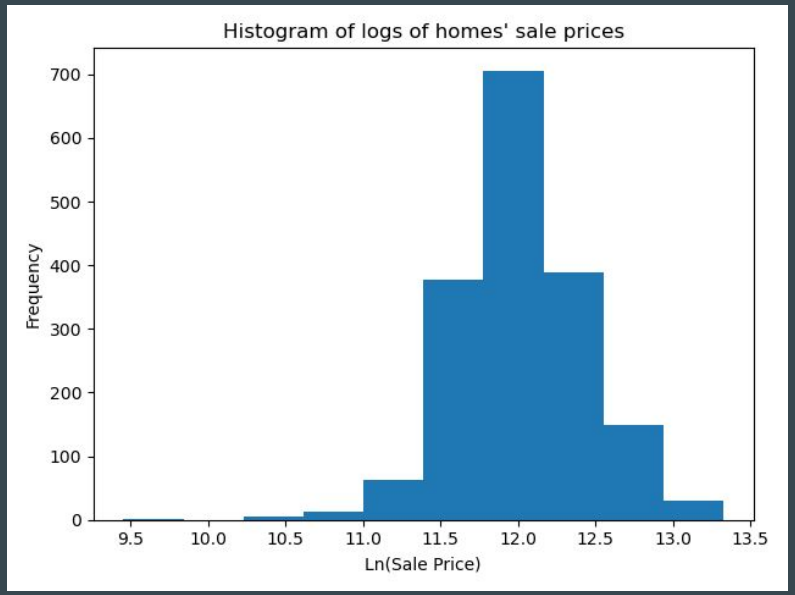
- Data: Ames Housing Dataset*
  - Contains data on over 75 features from over 2,000 homes sold in 2006-2010
- Models: Linear regression & regularized linear regression
- Evaluation:
  - Predictive power (on test data):
    - R^2 scores
    - Mean absolute error (MAE)
  - Interpretability: Can we infer effects of features on predicted sale prices?

---

* http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Summary of (Training) Data





Median: $163,000
Mean: $181,433
SD: $79,094

Baselines: Null Model and Optimal Regularized Models

- Null model scores on test data set:
  - R^2: 0
  - MAE: $60,525

- Regularized Regressions: LASSO, Ridge, ElasticNet

  - No interpretability!

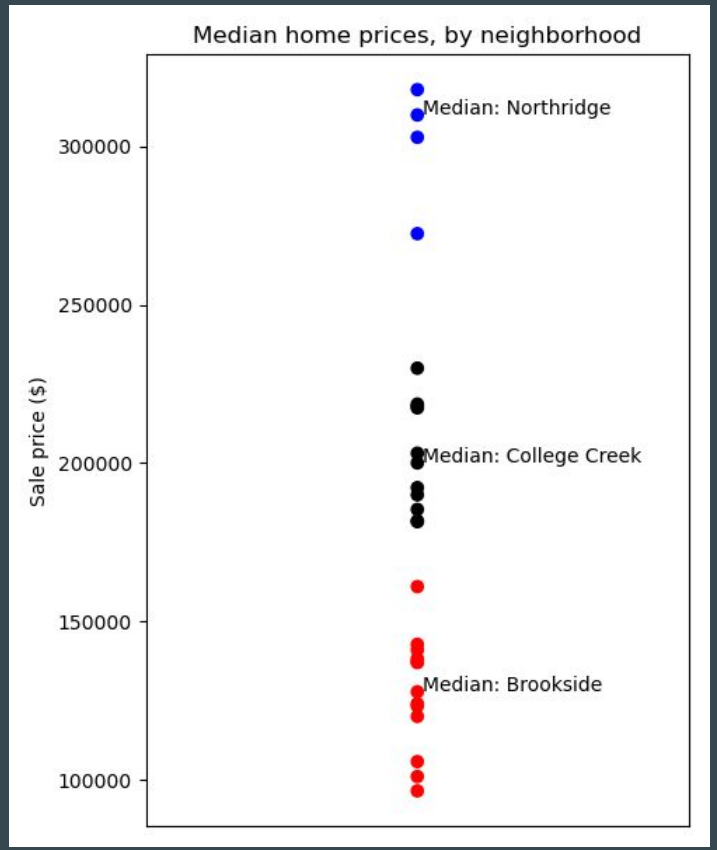
  - Best performer (gridsearch): LASSO predicting Ln(Sale Price). Test scores:
    - R2: .9286
    - MAE: $13,464

Interpretable Models

How to avoid overfitting but get good predictions?

How to get good predictions while keeping interpretability?

1. Encode categorical variables using above-below-mid (ABM) encoding. For Neighborhood, this looks like:



- 1-unit increase in Neighborhood means:
  - Moving from a "bad" to an "average" neighborhood, OR
  - Moving from an "average" to a "good" neighborhood

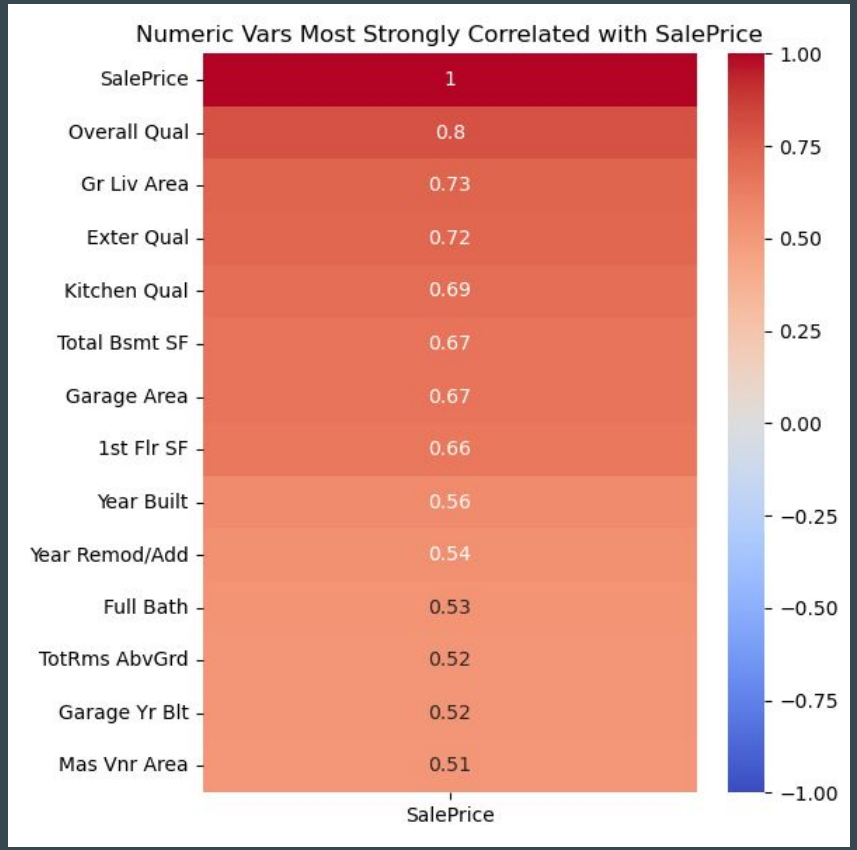
Northridge, Northridge Heights, Stone Brook, Veenker ("Good")

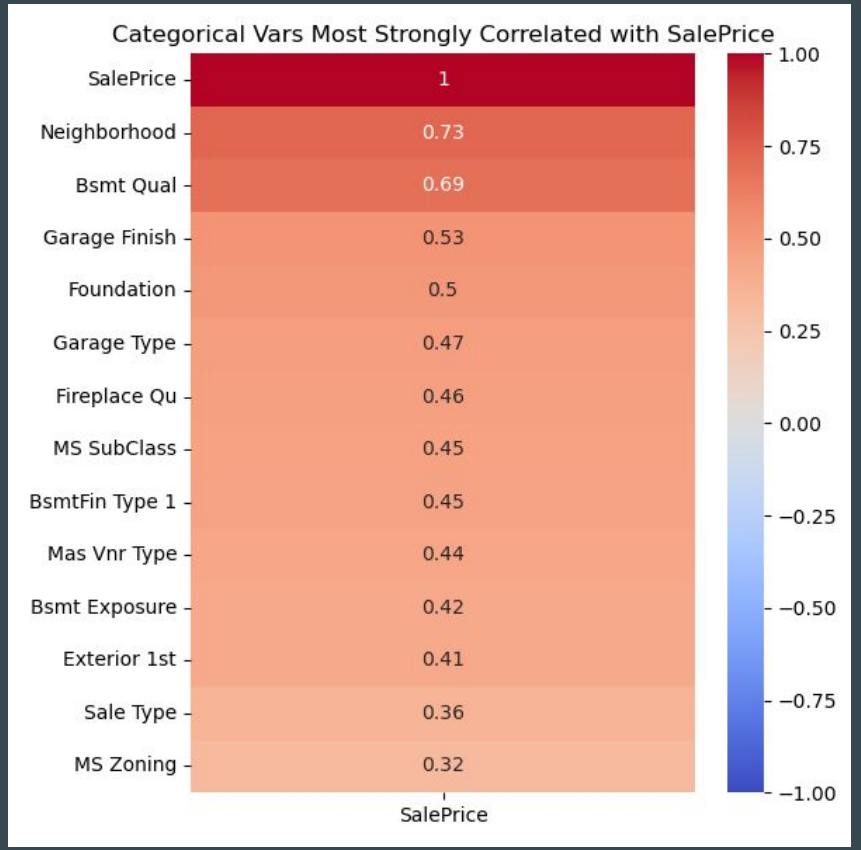
Bloomington Heights, Clear Creek, College Creek, Crawford, Gilbert, Greens, Green Hills, Northwest Ames, Sawyer West, Somerset, Timberland ("Average")

Bluestem, Briardale, Brookside, Edwards, Iowa DOT and Rail Road, Landmark, Meadow Village, Mitchell, North Ames, Northpark Villa, Old Town, South & West of Iowa State University, Sawyer ("Bad")

How to get good predictions while keeping interpretability?

2. Add features to a linear model, staring with those most highly correlated with sale price.



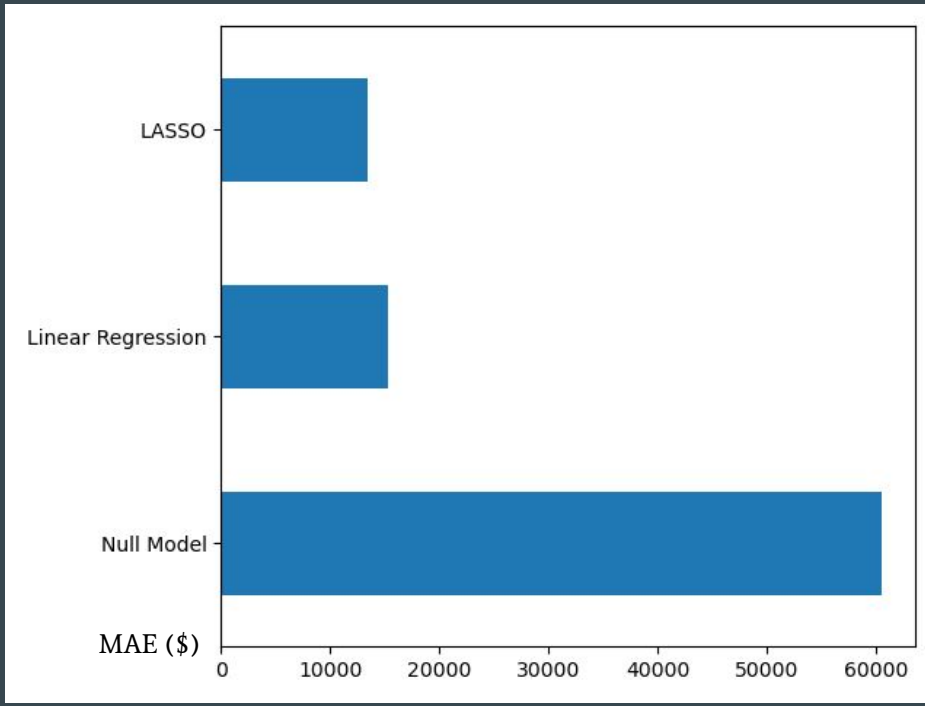


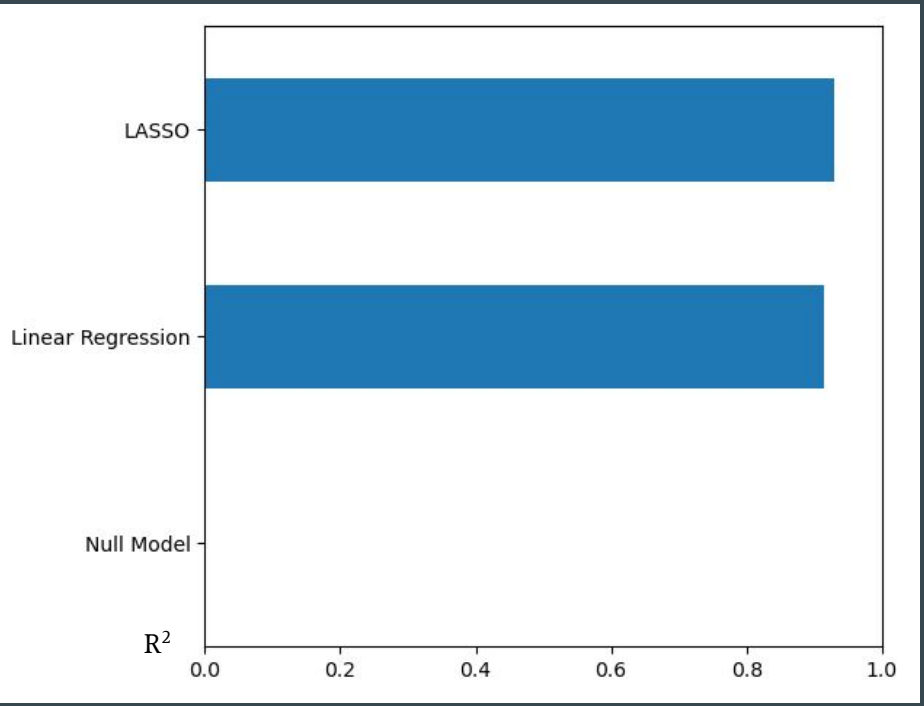
How to get good predictions while keeping interpretability?

3. Try all possibilities and select the model with best predictive power.

- Best linear regression predicting Ln(Sale Price):
    - 32 numeric features
    - 12 categorical features
    - Scores on test data:
        - R^2 = .915
        - MAE = $15,312

(Interpretable) Linear Model vs Null and LASSO

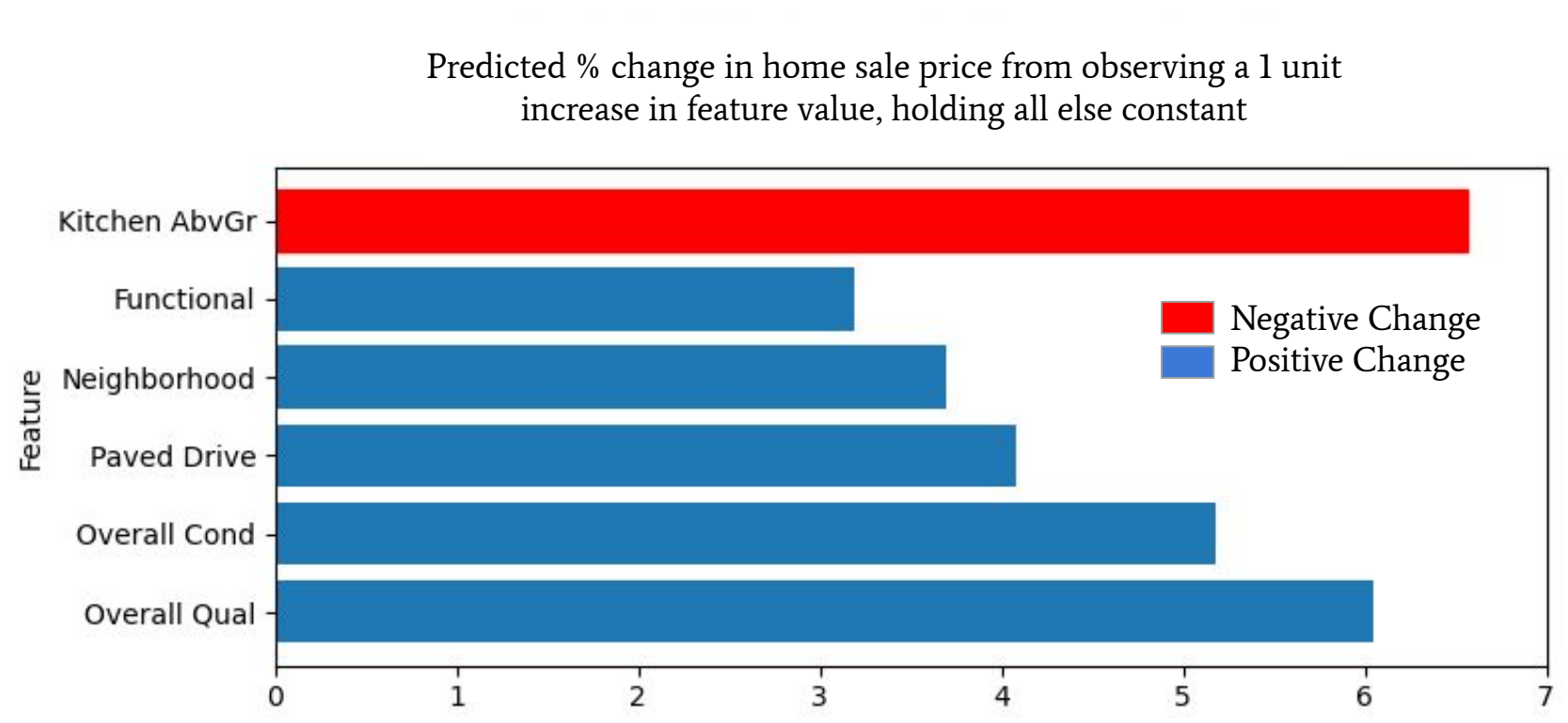




Interpretable model is worse than LASSO by only:

- $1,848 average prediction error
- 1.36% explanatory power

Interpretation of Best Linear Model



Conclusions

- A well-chosen linear regression can have high predictive power while maintaining interpretability

- Gains in predictive power from non-interpretable models may not be worthwhile

- Strongest predictors of higher home sale prices in Ames, IA (all else equal):
  - Quality of materials
  - Condition of home
  - Driveway pavement
  - Neighborhood