Ames housing dataset modelling challenge

Leonardo Blas

Can we predict housing prices?

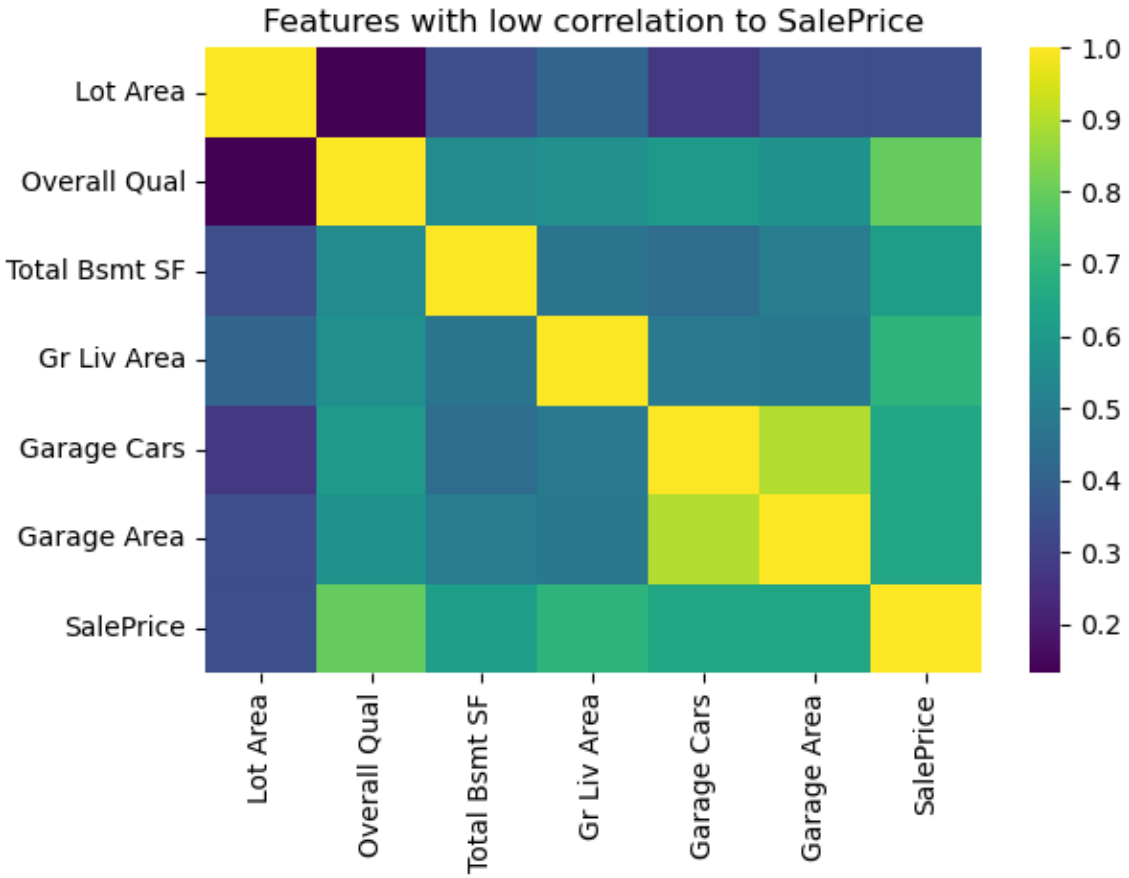
- We have:
  - A dataset of properties sold from 2006 to 2010 in Ames, IA.
  - ~1500 entries.
  - ~80 parameters.
- We want:
  - A model to predict housing prices.
  - To minimize the error in terms of dollars (RMSE).

Process overview

1. Consider the features we want to use.
2. Clean the data.
3. Transform the data (encode, standardize, normalize).
4. Try different models.
5. Choose the best performing model.

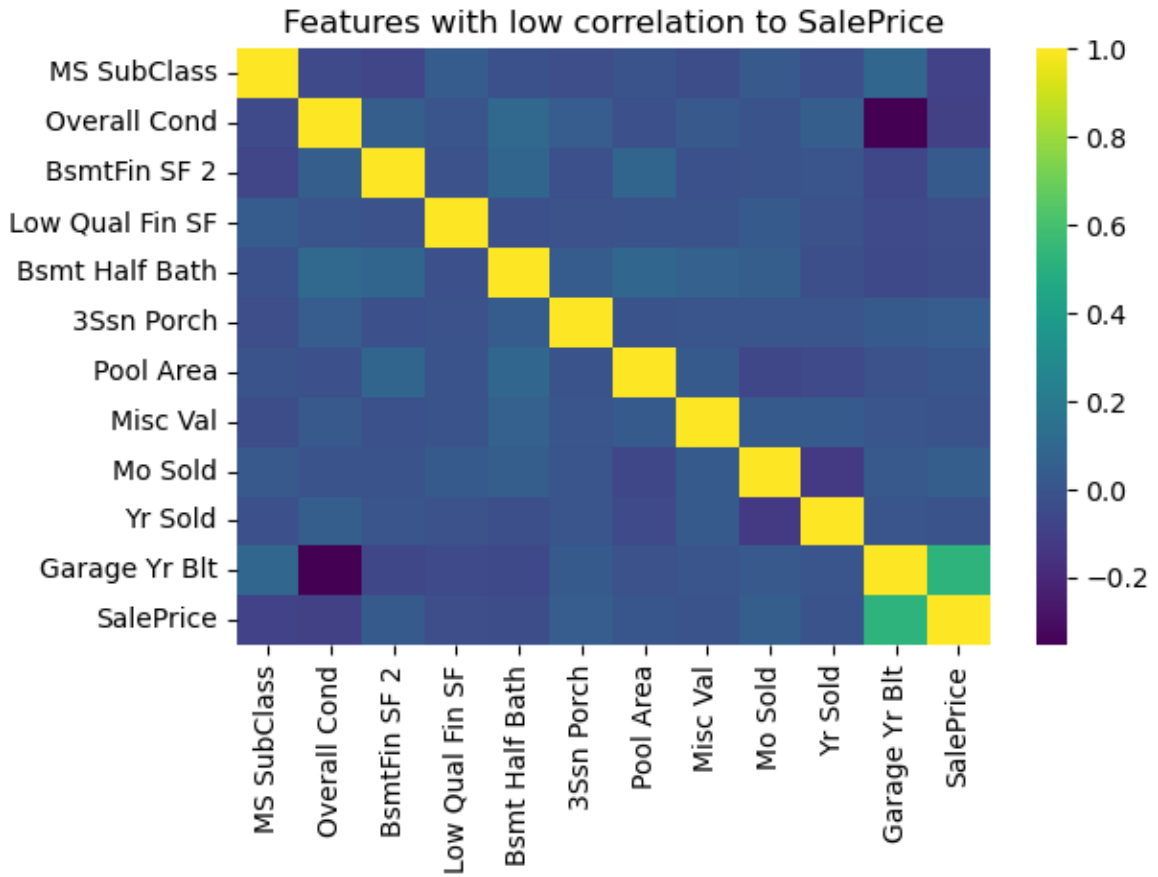
What features to use

- Overall quality.
- Living area area.
- Basement area.
- Garage car capacity.
- Garage area.



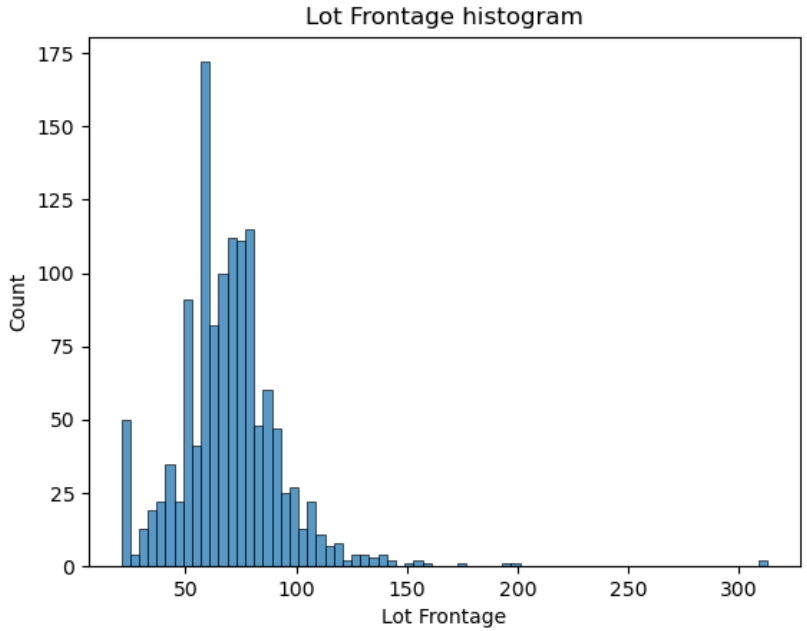
What not to use

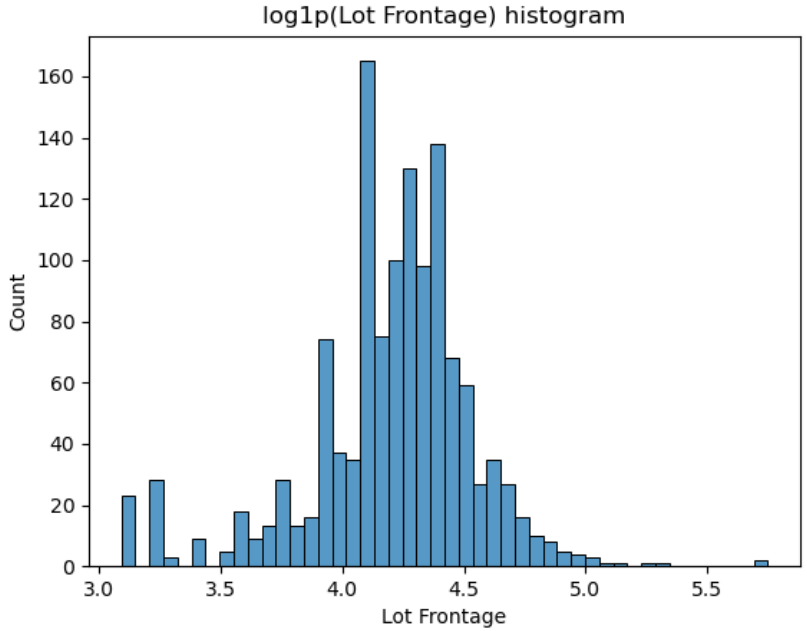
- Overall condition.
- Year sold.
- Month sold.
- Pool area.
- Basement half bathrooms.



Helpful transformations

- log1p.
- Power.
- KNN imputation.
- Standard scaling.

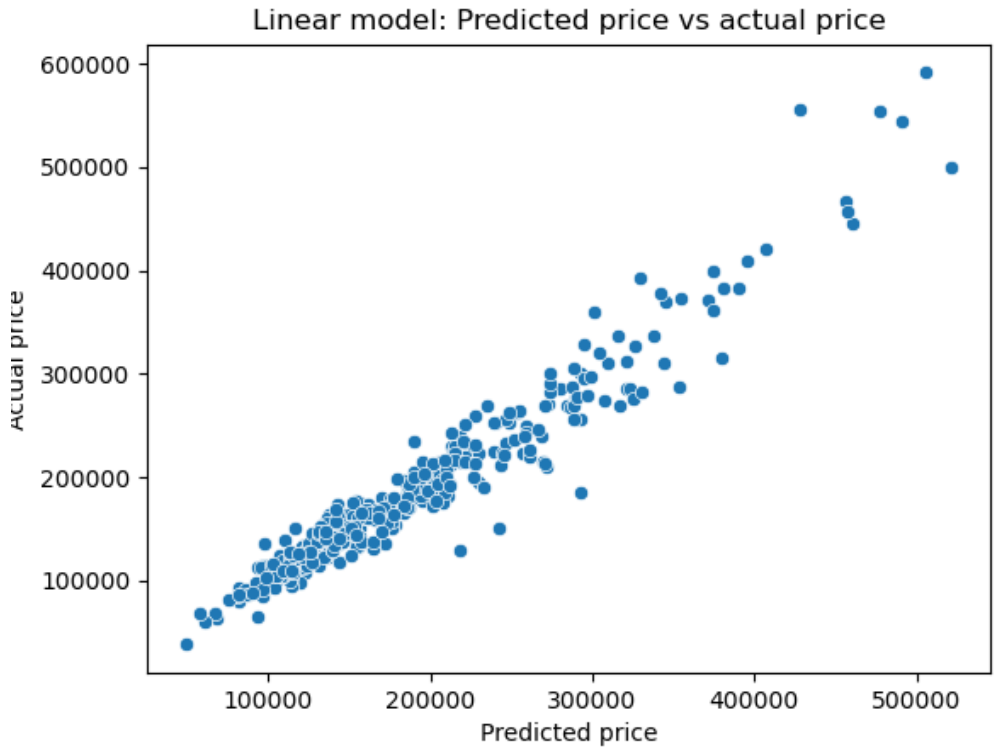




My models: Linear

- Train RMSE: ~$17k
- Test RMSE: ~$21k

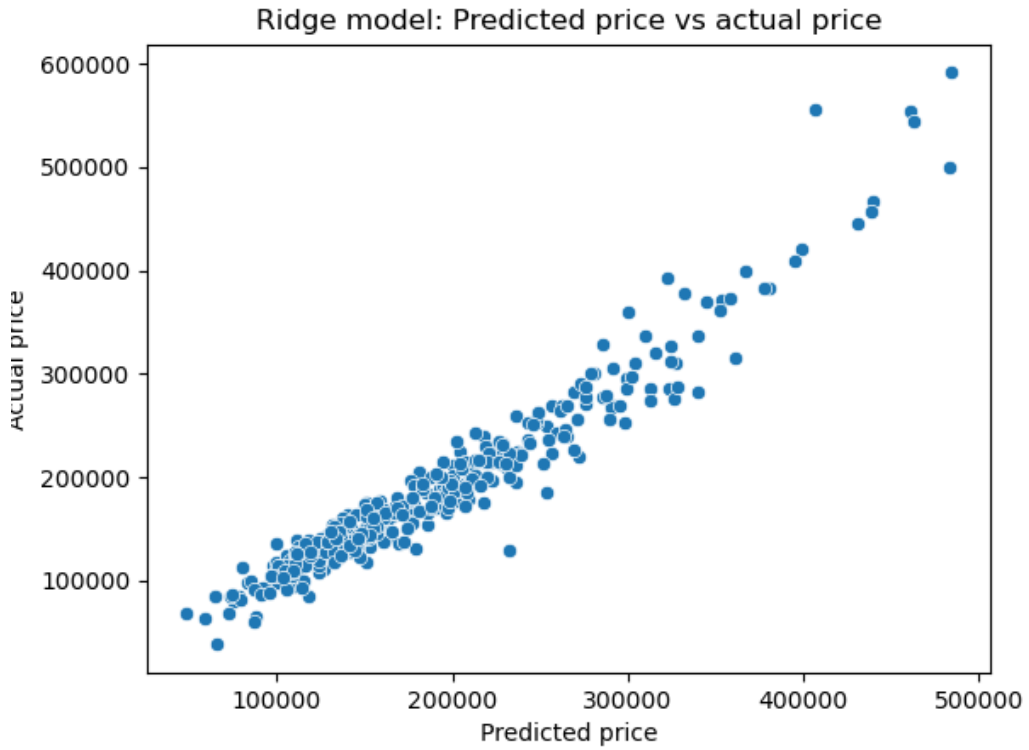
- Train r^2: ~0.95
- Test r^2: ~0.93



My models: Ridge

- Train MSE: ~$21k
- Test MSE: ~22k

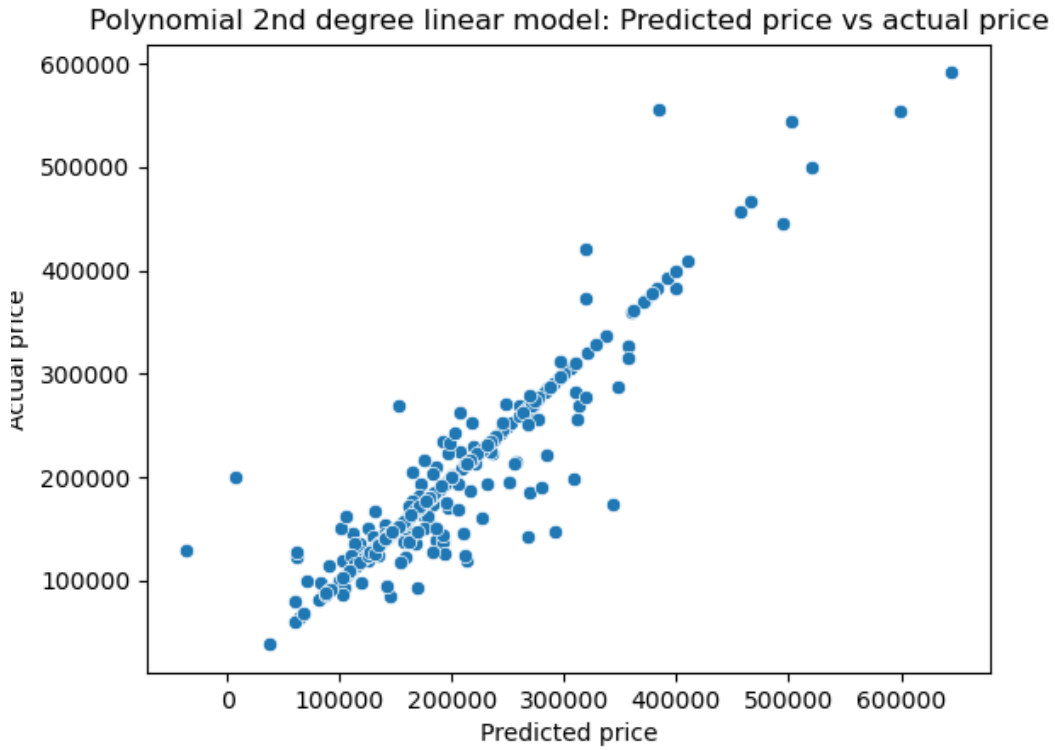
- Train r^2: 0.94
- Test r^2: 0.93



My models: Polynomial linear

- Train RMSE: ~$34k
- Test RMSE: ~$68k

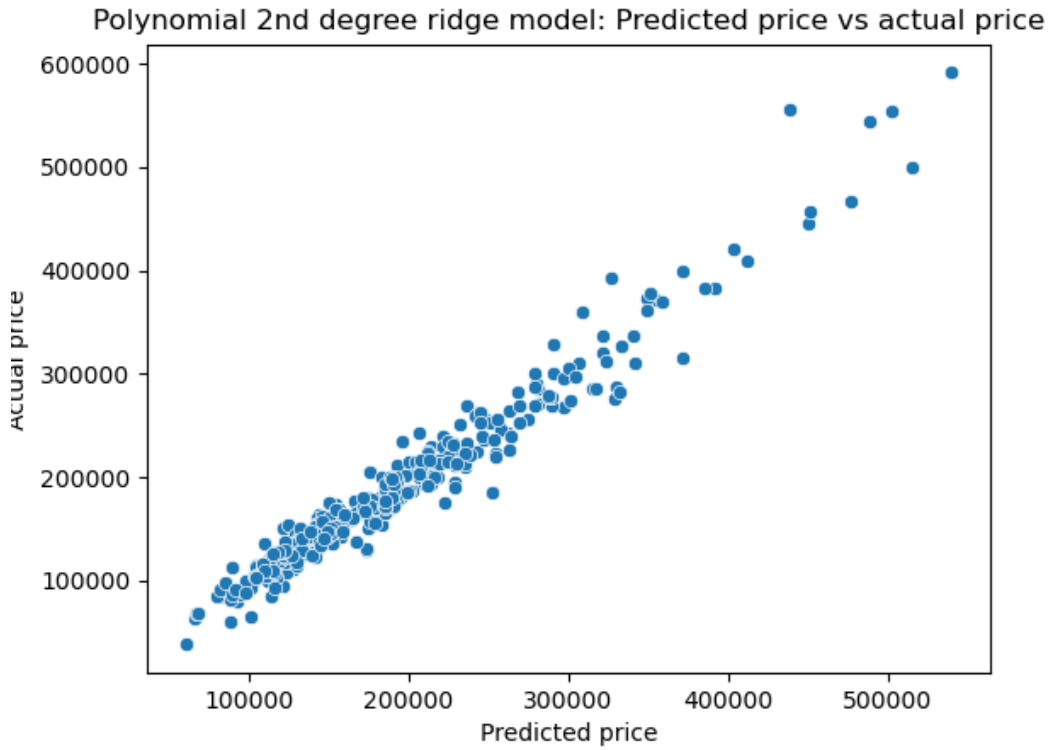
- Train r^2: 1.0
- Test r^2: 0.87



My models: Polynomial ridge

- Train RMSE: ~$19k
- Test RMSE: ~$17k

- Train r^2: ~0.97
- Test r^2: ~0.96



Findings and recommendations

- Data doesn’t lie: Build models considering the data patterns.
- It’s hard to summarize inter-parameter interactions.
- Parameters like overall quality and garage area are important.
- Some data is almost irrelevant.
- A 2^{nd} degree polynomial ridge model is efficient.

References

- [1] J. O, M. Harris, "Ames Iowa Submission". 2024. https://kaggle.com/competitions/adobe-dsb-34