Predicting the Sale Price of Residential Properties

Joshua Ogden-Davis

Springboard Data Science Track

June 2023 Cohort

Capstone Project

Problem

Purchasing a home is not only one of the most important and consequential financial decisions in American life, it is a crucial part of the economy. Residential fixed investment and housing services together account for roughly 15-16% of the US GDP in 2023.*

However, both the official valuation of a residential property as well as the sale price negotiation process are subject to biases and sales tactics that can make navigating this decision treacherous and potentially unequitable, with profound consequences for individuals, families, and the economy as a whole.

Our task is to develop a machine learning model that will reliably estimate the expected sale price for residential properties based on objective features as a useful tool to reduce uncertainty for homebuyers and investors.

* https://eyeonhousing.org/2023/10/housing-share-of-gdp-remains-flat-in-the-third-quarter-of-2023/

What might affect a property's price?

Features	Location	Condition
House size	Neighborhood	Year built
Lot size	Zoning	Year remodeled
Number of rooms	Nearby amenities	Exterior quality
Materials	Highway access	Interior quality
Style	Alley access	Finished interior %
(etc)	(etc)	(etc)

Data Source

The Ames Housing Dataset was compiled in 2011 by Dean De Cock from residential property sale information obtained through the Ames City Assessor's Office regarding sales that took place between 2006 and 2010. It contains 79 features for 2,930 records, each record pertaining to a sale.

I will be using a subset of the Ames dataset that has been pre-divided into a training set (1,460 records) and a test set (1,459 records) by Kaggle.

* De Cock's paper about the dataset can be found here: https://jse.amstat.org/v19n3/decock.pdf

Data Preparation

The features fall into one of three categories, requiring different preparation:

Continuous

Sizes (in square feet)
Quantities (ie, bedrooms)

- Remove unreasonable values
- Remove outliers (>3.5std)
- Impute missing values
- Normalize

Ordinal

Quality scores

- Convert from strings to integers
- Impute missing values
- Normalize

Categorical

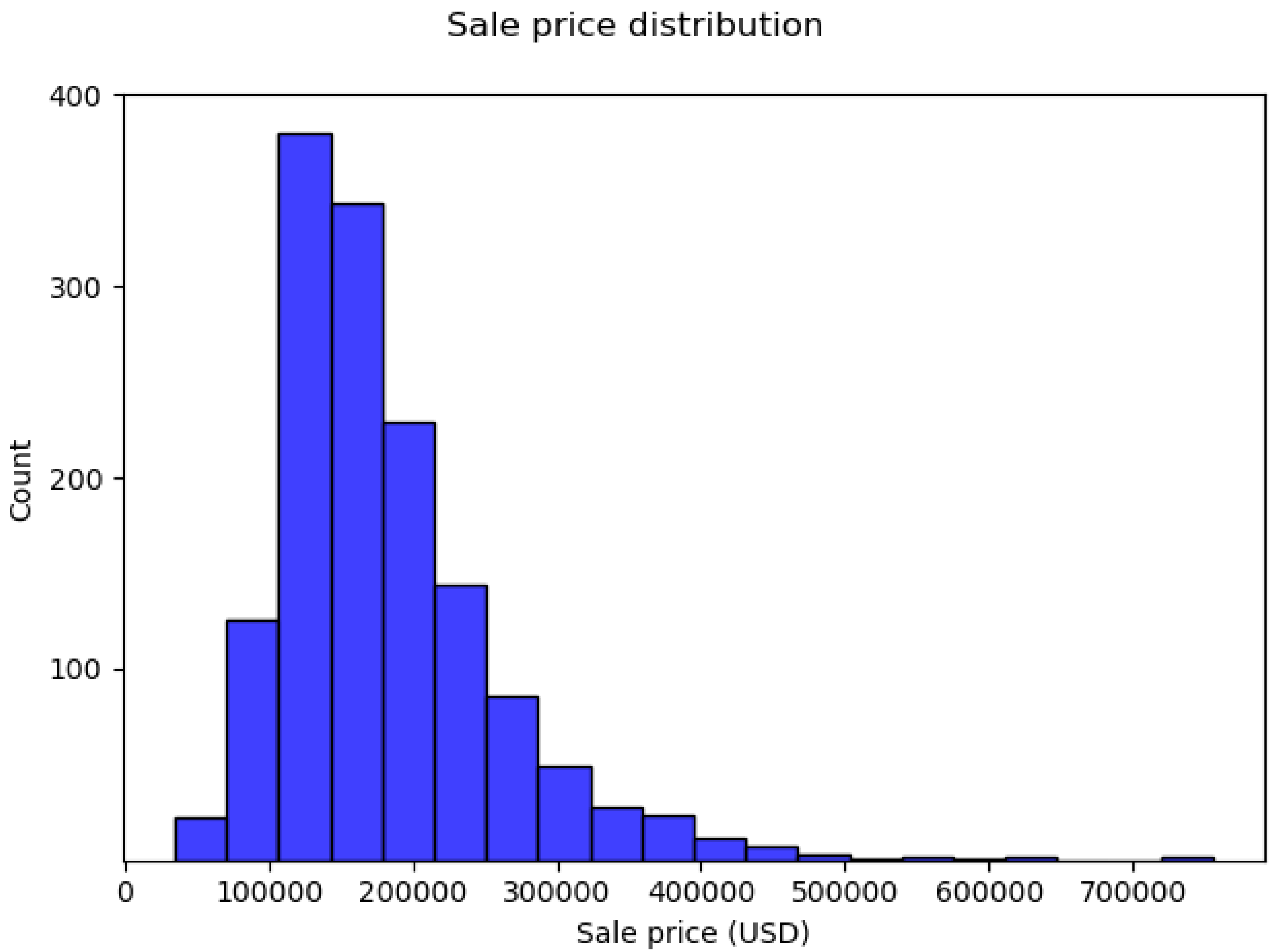
Neighborhood
Materials (wood, brick, etc)
Etc.

- Standardize spellings
- Impute missing values
- Check for importance
- One-hot encode

Target variable

- Right skew with outliers (expected)

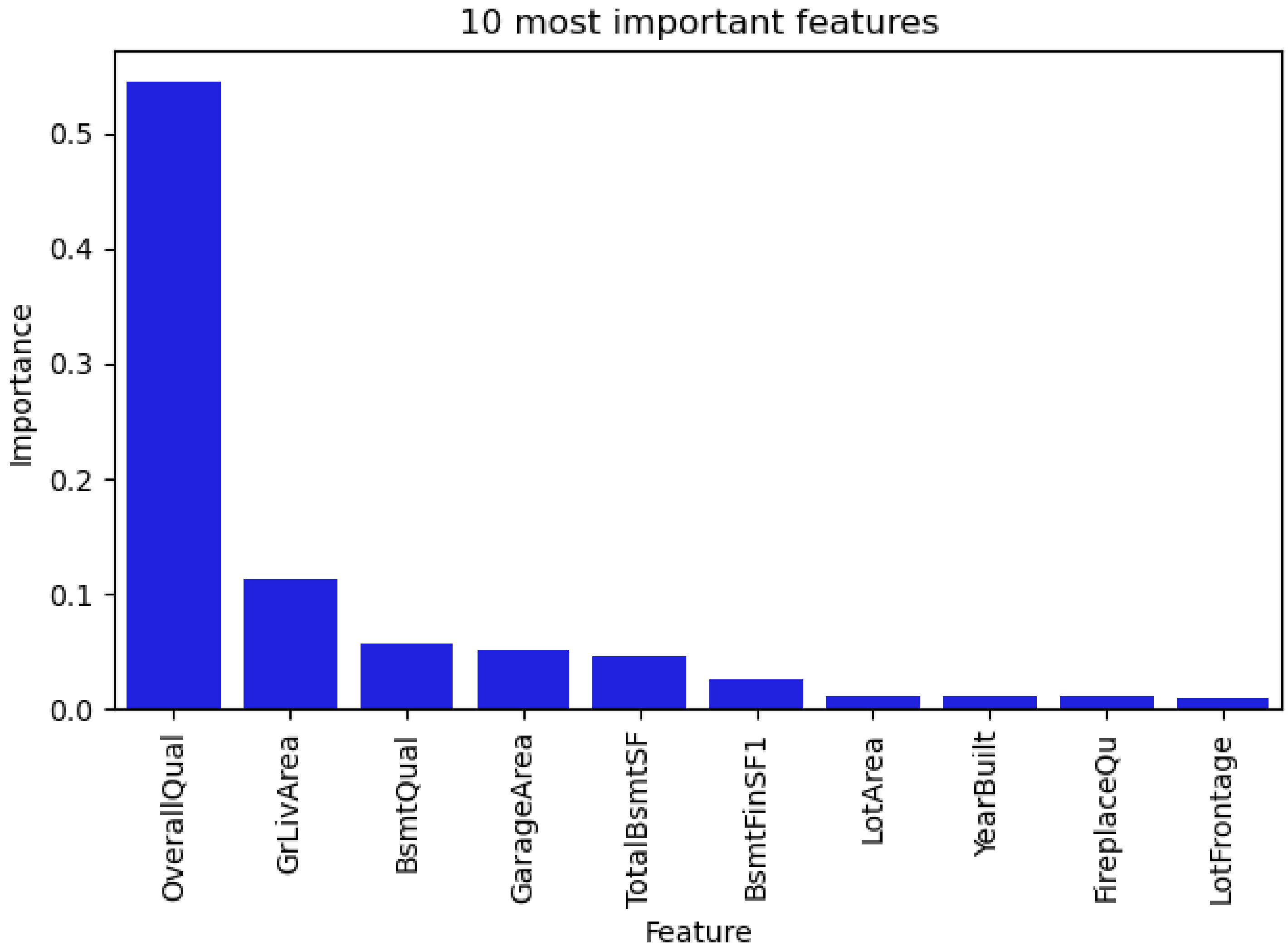
- No left outliers



Feature importances

- Quality is by far the most important

- Most important features are quality scores and size measurements

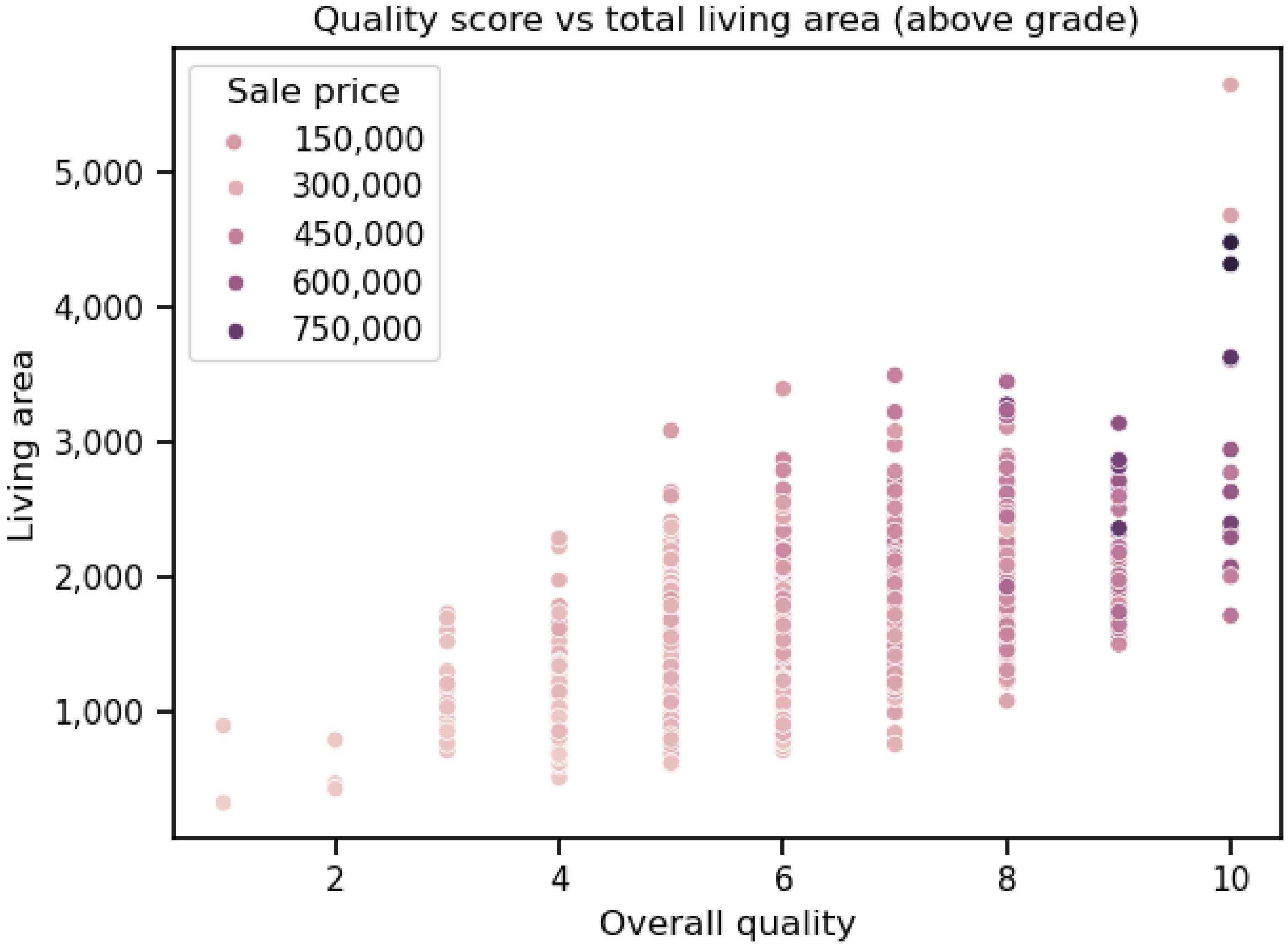


Feature surprises: overall quality vs area

Our two most important features have an interesting correlation.

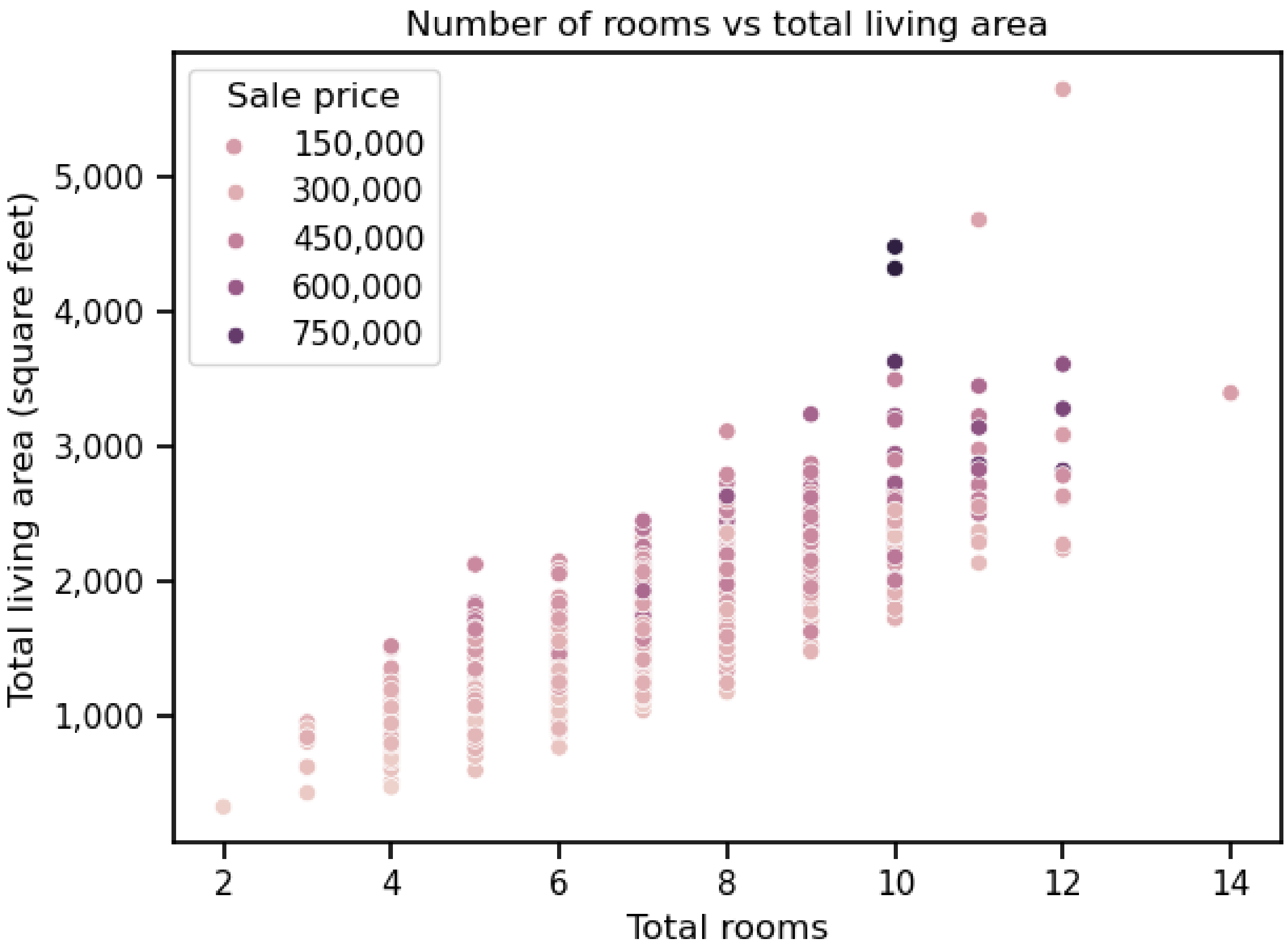
Homes above a certain size have narrower quality score ranges, until 10 is reached.

This suggests that evaluators score 10s differently than other categories; it may be more likely that they lump a house into the 10 category, instead of 8 or 9.



Feature surprises: rooms vs area

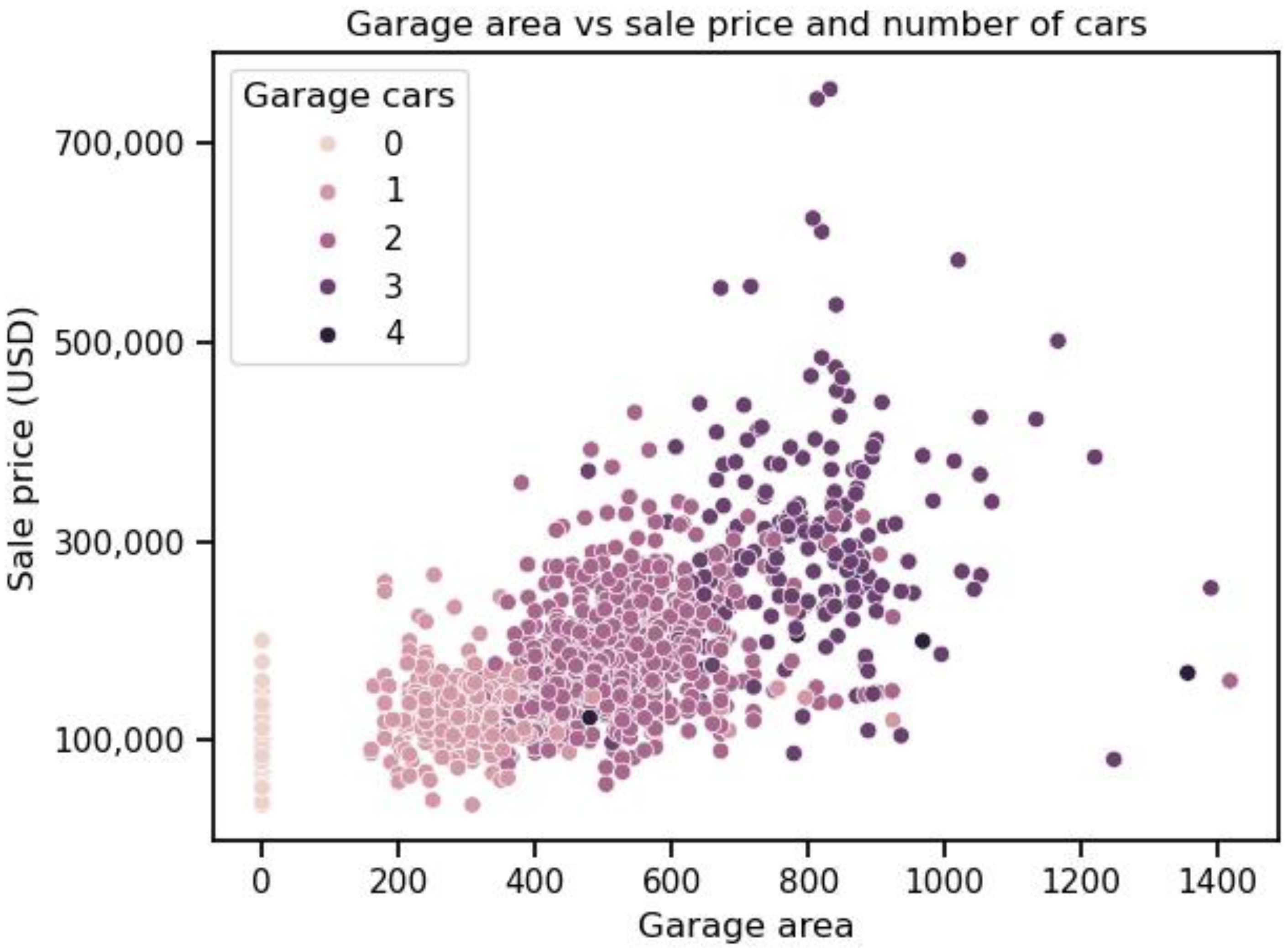
While size matters, there is a point of diminishing returns for both size and number of rooms.



Feature surprises: garage area vs type

While garage size greatly influences sale price, there's a wide range of sizes for each type of garage.

Some figures (such as the 500sqft, 4-car garage) seem improbable and are likely errors.

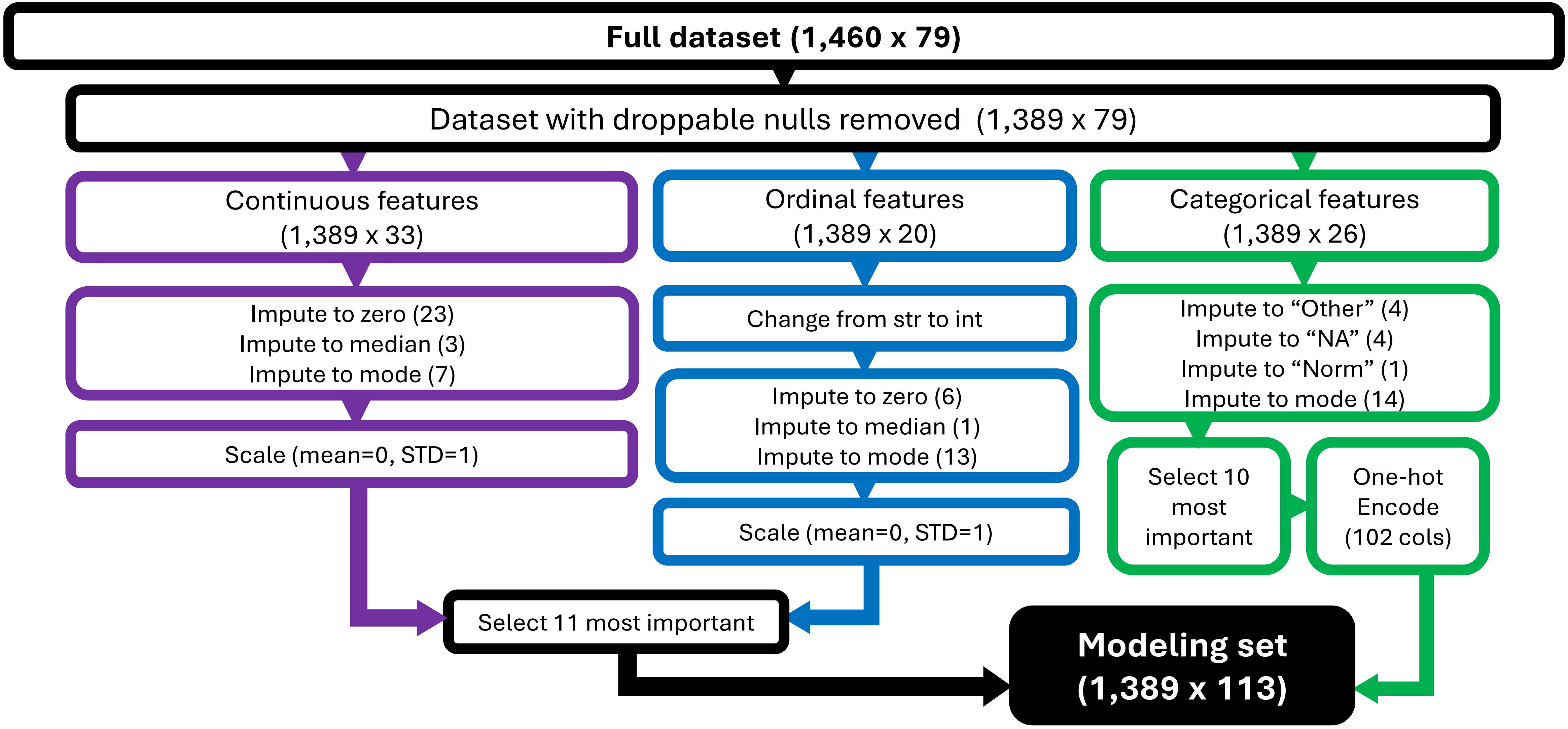


Feature surprise takeaways

The variables have generally predictable but imperfect and occasionally surprising relationships.

This suggests an **ensemble model**, such as **gradient boost**, may be best able to capture the points-of-no-return and rating biases.

Pre-Preparation Process

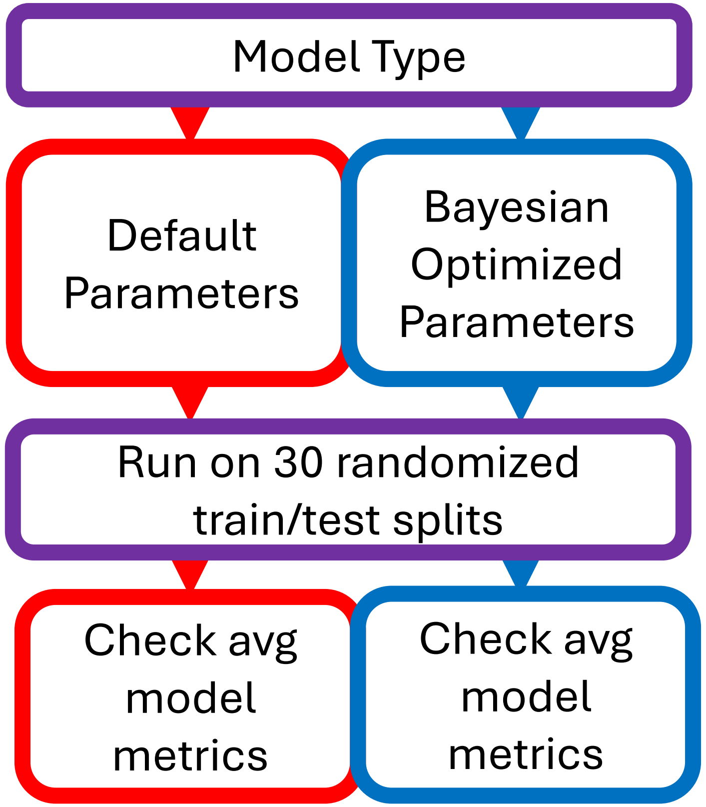


Modeling process

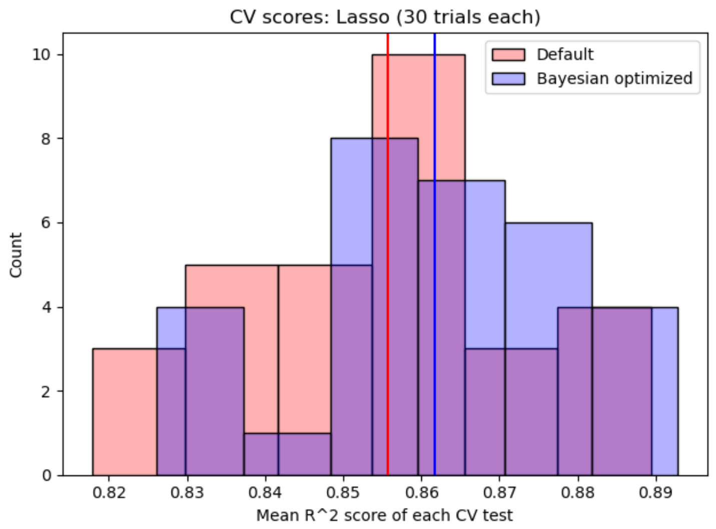
- Given the nature of the problem described earlier, I expected gradient boosting or random forest to yield the best results.

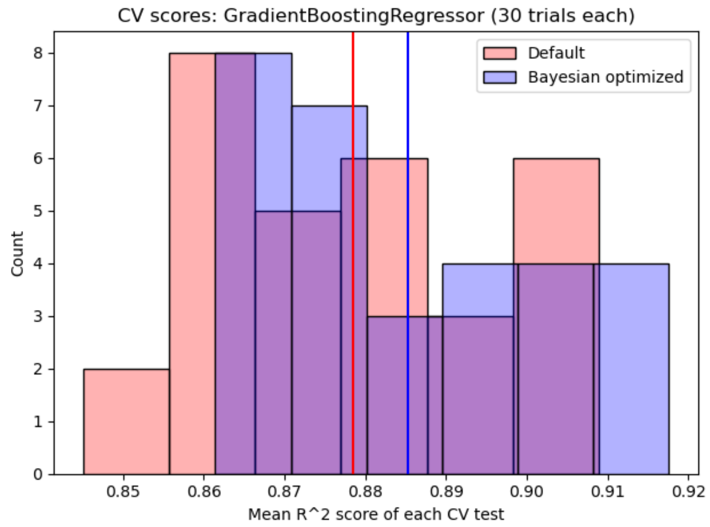
- I also included a Lasso model as a sanity check.

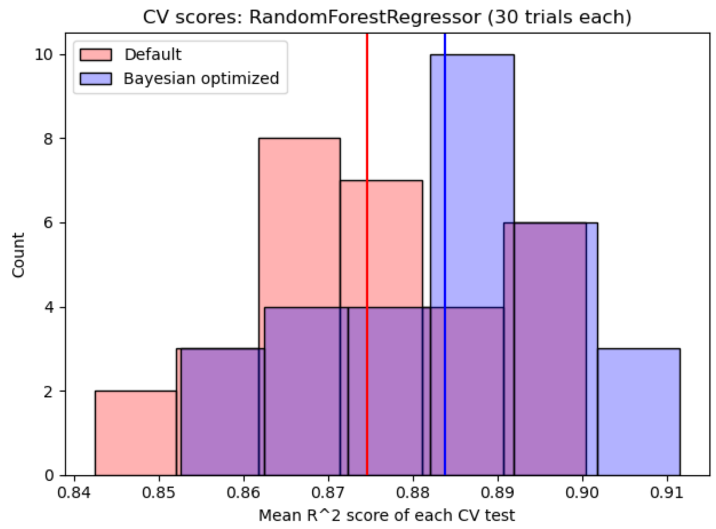
- Each model went through the following process:



Model type performances







Mean Lasso R^2 score: .866

Mean GB R^2 score: .885

Mean RF R^2 score: .884

Since the mean GB and RF scores were comparable, I ran a t-test to see if the difference was possibly due to chance and got a p-value of 0.72.

Thus, I am not confident that GB or RF is consistently more accurate than the other.

Winning model parameters and metrics (GB)

Optimal parameters (Bayesian optimized):

'learning_rate':	0.0249
'max_depth':	10
'min_samples_leaf':	5
'min_samples_split':	12
'n_estimators':	291
'subsample':	0.2108

Performance metrics:

R^2 score:	0.885
Explained variance:	0.885
Mean squared log error:	0.020
Mean squared error (USD^2):	452,431,840.99
Mean absolute error (USD):	15,260.78

Limitations

The model’s applicability to real-life situations suffers from the following:

- Generalizability
The data is limited to a small location that is not representative of the country.

- Relevance
The data comes from a narrow range of years over a decade ago.

- Reliability
The data contains some clear errors, and the collection process is not transparent.

Future improvements

- **Improved feature engineering**
  More in-depth exploration of feature selection and creation could yield better results.

- **Expert consultation**
  Specific domain expertise could offer useful insights.

- **Model combination**
  Employing different models and combining the results may reduce error.

Thank you!

Joshua Ogden-Davis

Springboard Data Science Track

June 2023 Cohort

Capstone Project