Predicting market value of single family homes in Philadelphia

Philadelphia real estate

Property ownership is an attractive brick-and-mortar, long-term investment[1]

Increase in net migration from New York to Philadelphia in recent years

attributed to job growth and affordability [2]

What drives market value of single family homes in Philadelphia?

1. Rao, Krishna. (2014, May 8). Is a Home a Good Investment? Zillow Research. https://www.zillow.com/research/returns-to-housing-6874/

2. Central Philadelphia Development Corporation (2020, September 17). More New Yorkers Moving to Philadelphia: Job Growth Plus Affordability Key Attractors. https://www.centercityphila.org/cpdc/cpdc-news/more-new-yorkers-moving-to-philadelphia-job-growth

Philadelphia real estate

The Office of Property Assessment offers public records of individual properties in the city, including home features and estimated market value [3]

used to estimate property (real-estate) tax [4]

goal: 1) predict market value of single family homes in Philadelphia and
2) determine which features of homes are the best predictors of market value

could be of interest to current or potential future home investors and individuals who pay property tax

3. Data source: https://www.phila.gov/property/data/#

4. Office of Property Assessment, City of Philadelphia. “What we do.” https://www.phila.gov/departments/office-of-property-assessment/

Philadelphia real estate

scope of prediction: focusing on homes that may interest in the average home buyer (excluding very expensive or inexpensive properties)

approach: develop regression models to predict market value, with emphasis on those with higher interpretability [tree-based]

Variables included in dataset

market value
location (latitude, longitude)
zip code
depth of home structure
width (frontage) of home structure
total livable area
total area
garage space
number of bedrooms
number of bathrooms
view
year built
basement type
central air
number of stories
type of heater
interior condition (assessed by OPA)
exterior condition (assessed by OPA)

downloaded public record of property assessments from
https://www.phila.gov/property/data/#

downloaded metadata (info on variables) from
https://metadata.phila.gov/#home/datasetdetails/5543865f20583086178c4ee5/representationdetails/55d624fdad35c7e854cb21a4/

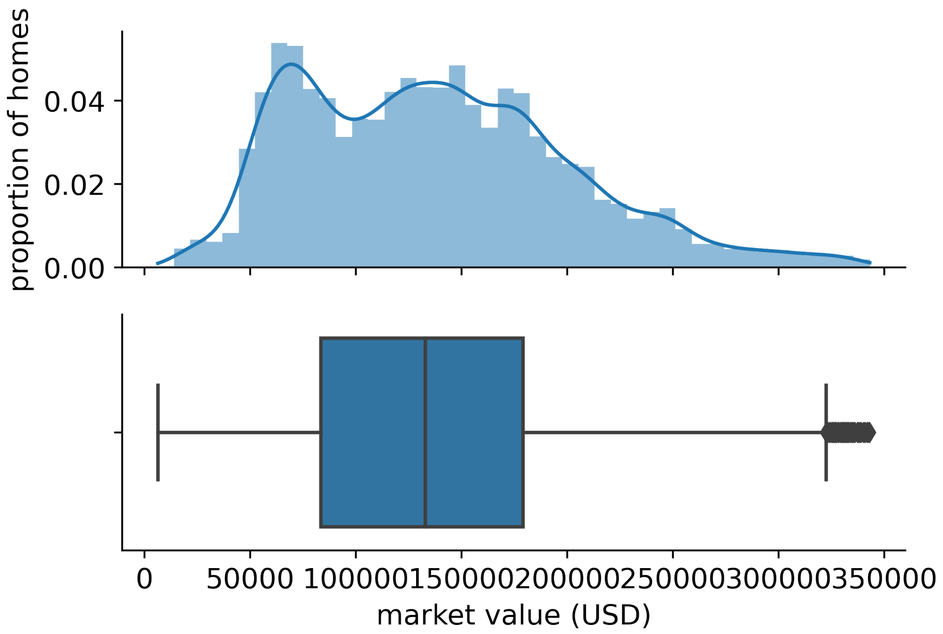
added distance to schools to this dataset
—calculated Manhattan distance between each home and the nearest elementary, middle, and high school

downloaded school locations info (type, latitude, longitude) in a csv file from

https://www.opendataphilly.org/dataset/schools/resource/8e1bb3e6-7fb5-4018-95f8-63b3fc420557

How is market value distributed for single family homes (excluding outliers)?

distribution is based on a filtered dataset, excluding outliers (+/- 1.5 iqr)

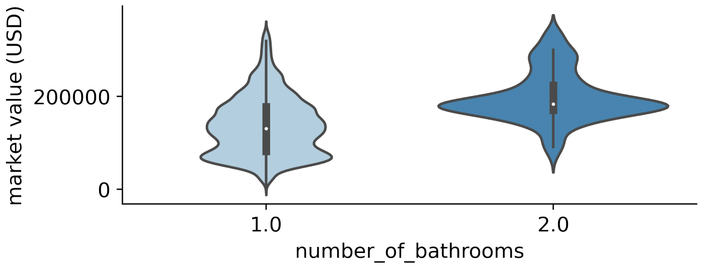


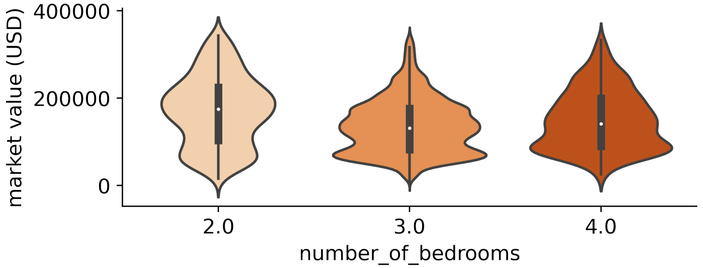
For this project, the middle 50% of homes fall within a $95,600 range

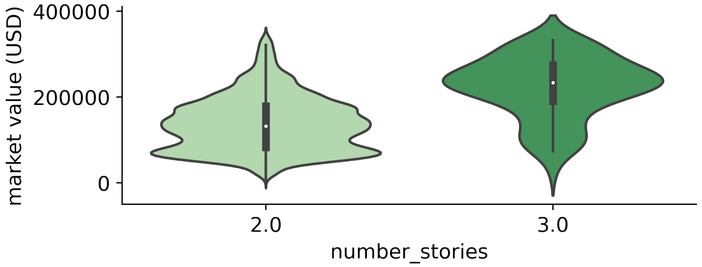
The variance in the upper 25% was greater than the lower 25%

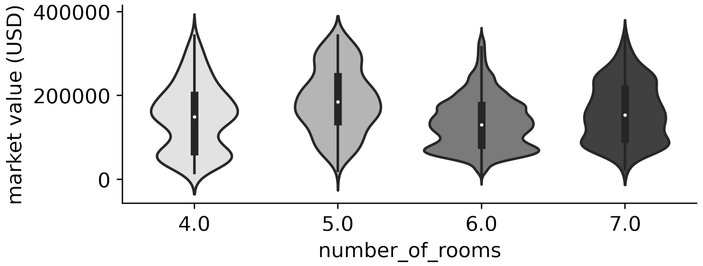
The upper 25% fell in a $164,100 range.

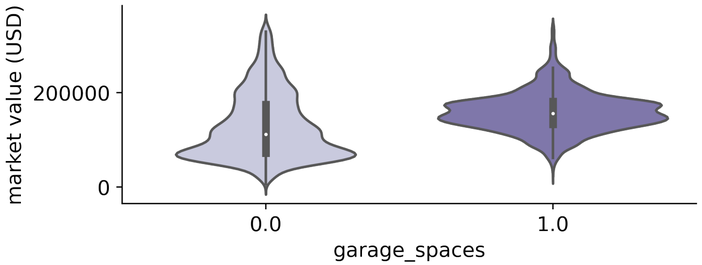
How does market value vary by number of rooms, stories, garage space?





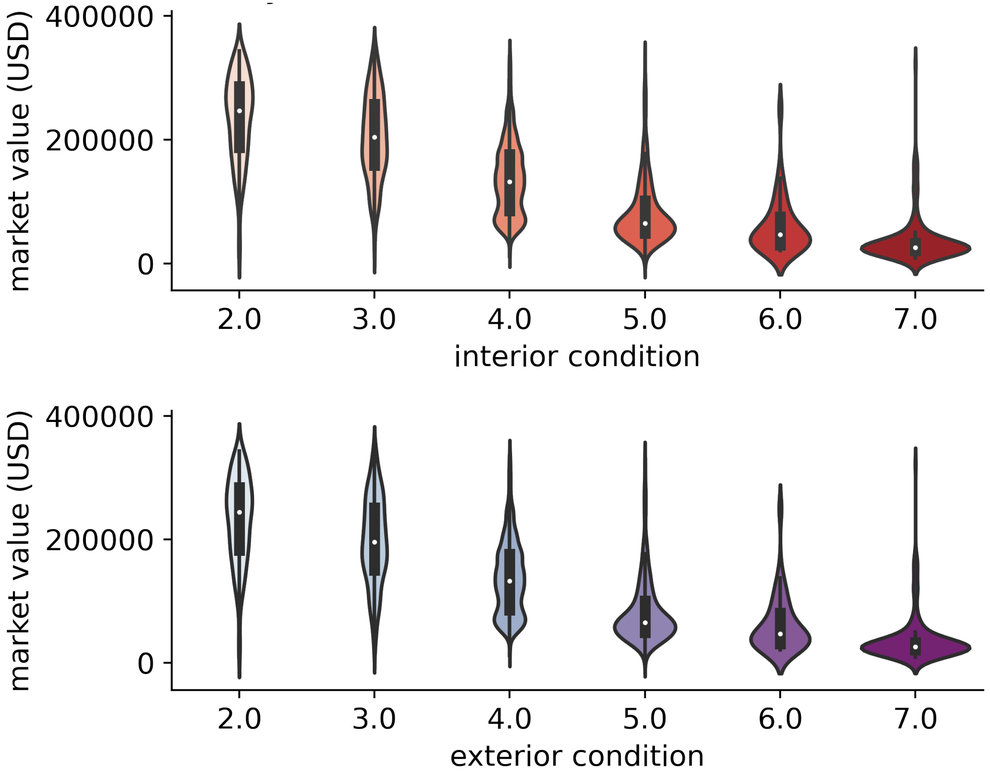






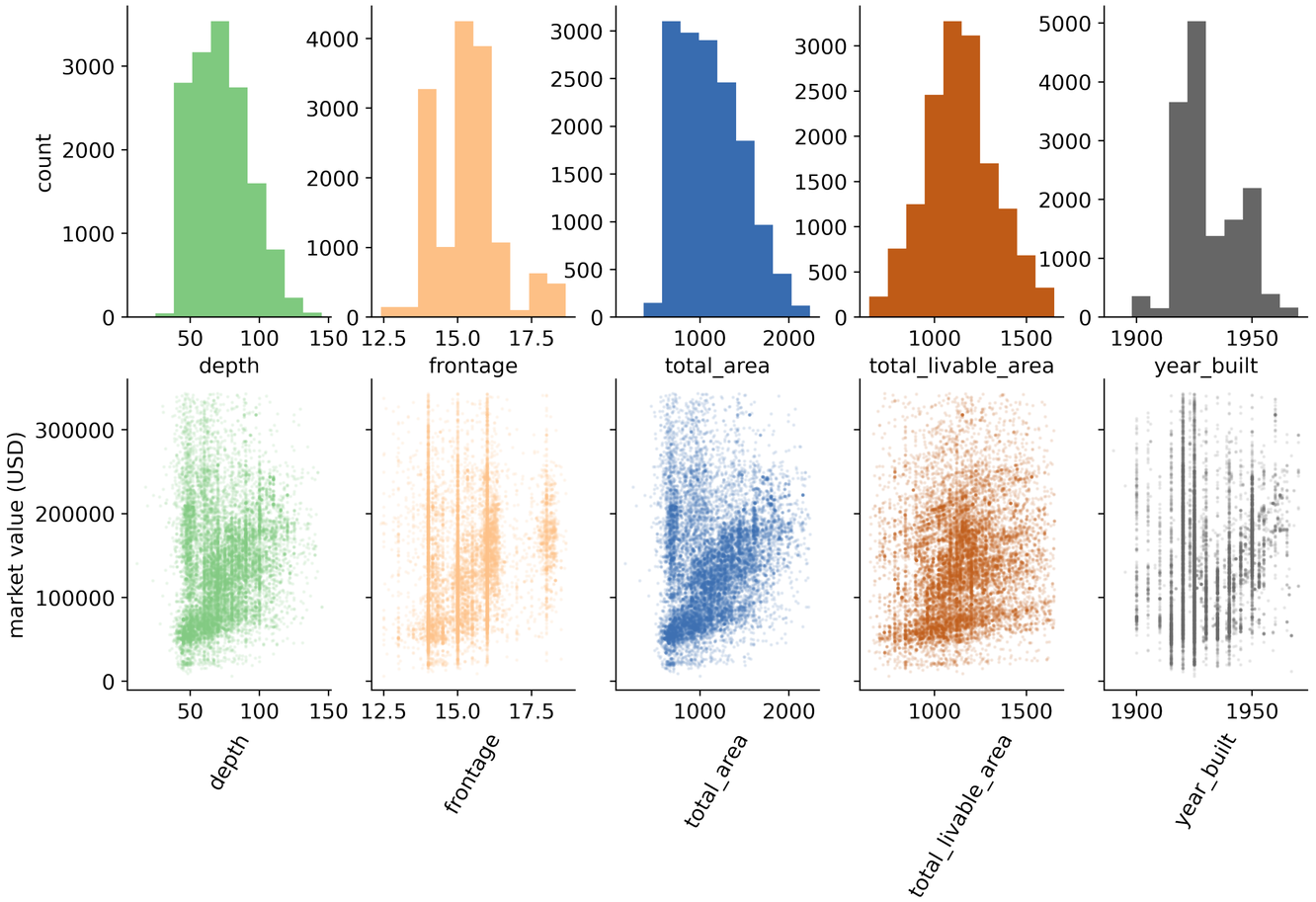
Market values are higher for more bathrooms, stories, and garage space

How does market value vary by home condition?



Market values are higher for higher interior and exterior condition grades

How does market value with home dimensions and age?



Market values are generally higher for larger homes

Regression models trained and evaluated

1. Linear Regression

to compare more advanced models to a basic model

2. Decision Tree Regressor
3. Gradient Boosting Regressor
4. Random Forest Regressor
5. Light Gradient Boosting Regressor

interpretable tree-based models
-may be useful in determining features that drive up market value

Regression models trained and evaluated

1. Linear Regression
   scaled data
2. Decision Tree Regressor
   raw data
3. Gradient Boosting Regressor
   raw data
4. Random Forest Regressor
   raw data
5. Light Gradient Boosting Regressor
   raw data

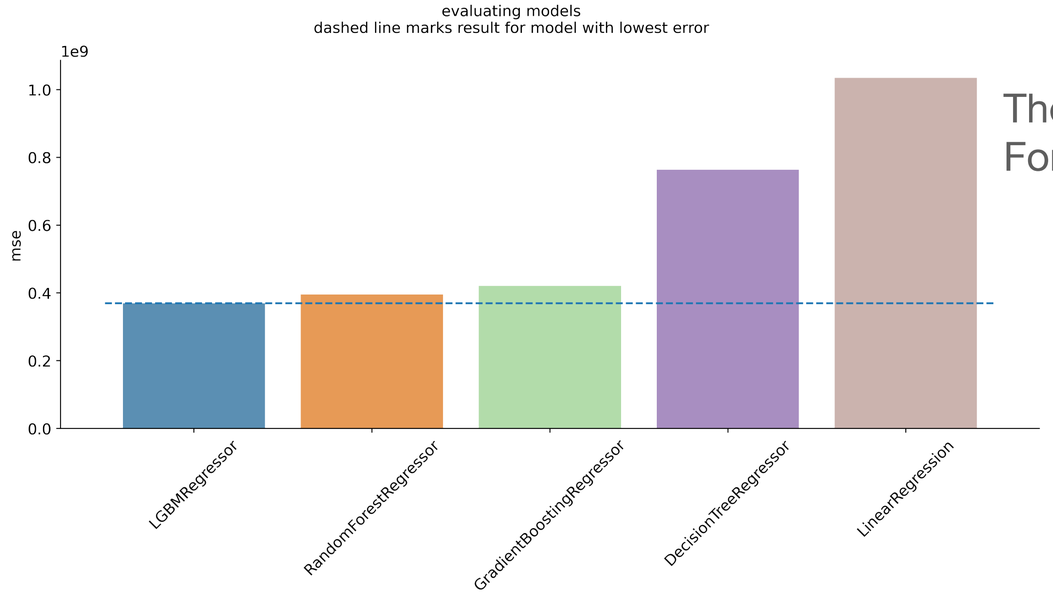
split data into training and test sets (proportion = 0.3)

ran 5-fold cross-validation (CV) to tune hyperparameters

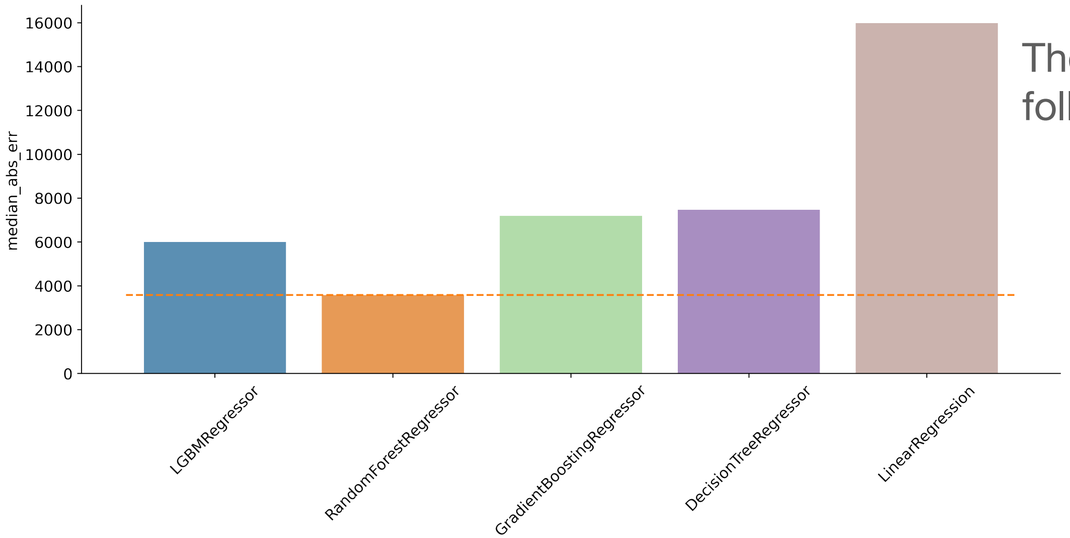
predicted test data using best estimator

assessed models using mean square error, median absolute error as well as the variance across CV folds

Market value by predicted market value for tested models



The LGBM Regressor has the lowest MSE, followed by the Random Forest Regressor

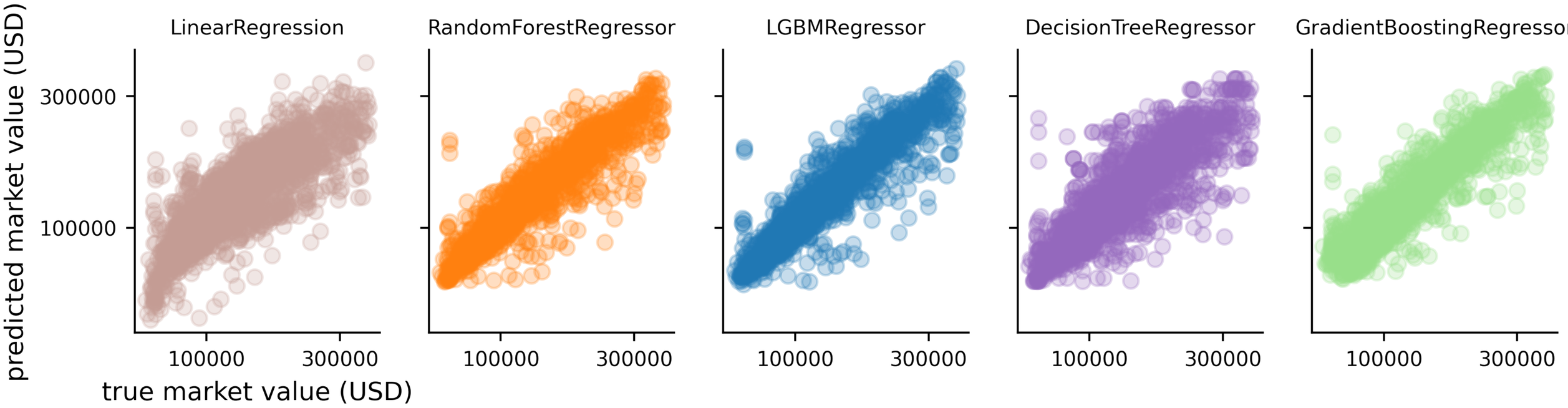


The Random Forest Regressor has the lowest median absolute error, followed by the LGBM Regressor

-Both models had similar variance in scores across CV folds

Chose LGBM Regressor because dataset is already reduced to exclude outliers (MSE may be fine) and this model trained a magnitude of time faster than the other

Actual market value by predicted market value for tested models



Visual inspection reveals greater dispersion away from unity line for Linear Regression, Decision Tree Regressor

Compared train LGBM Regressor model to 100 random, shuffled models

1.Took same training data from ‘best’ model

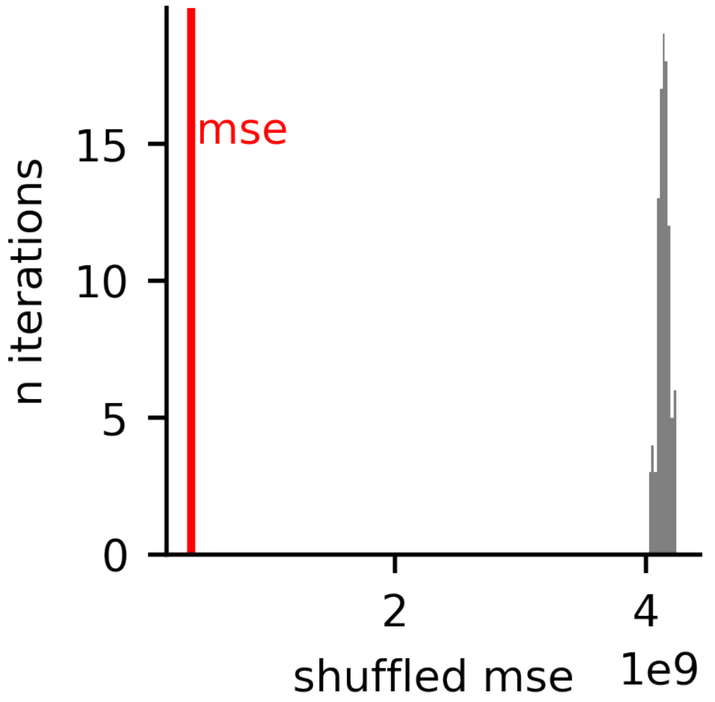
-kept X data (rows) intact
-shuffled y (target: market value) values

2. Trained model as before (5-fold cross-validation)

3. Predicted market value of X test data using trained shuffled model

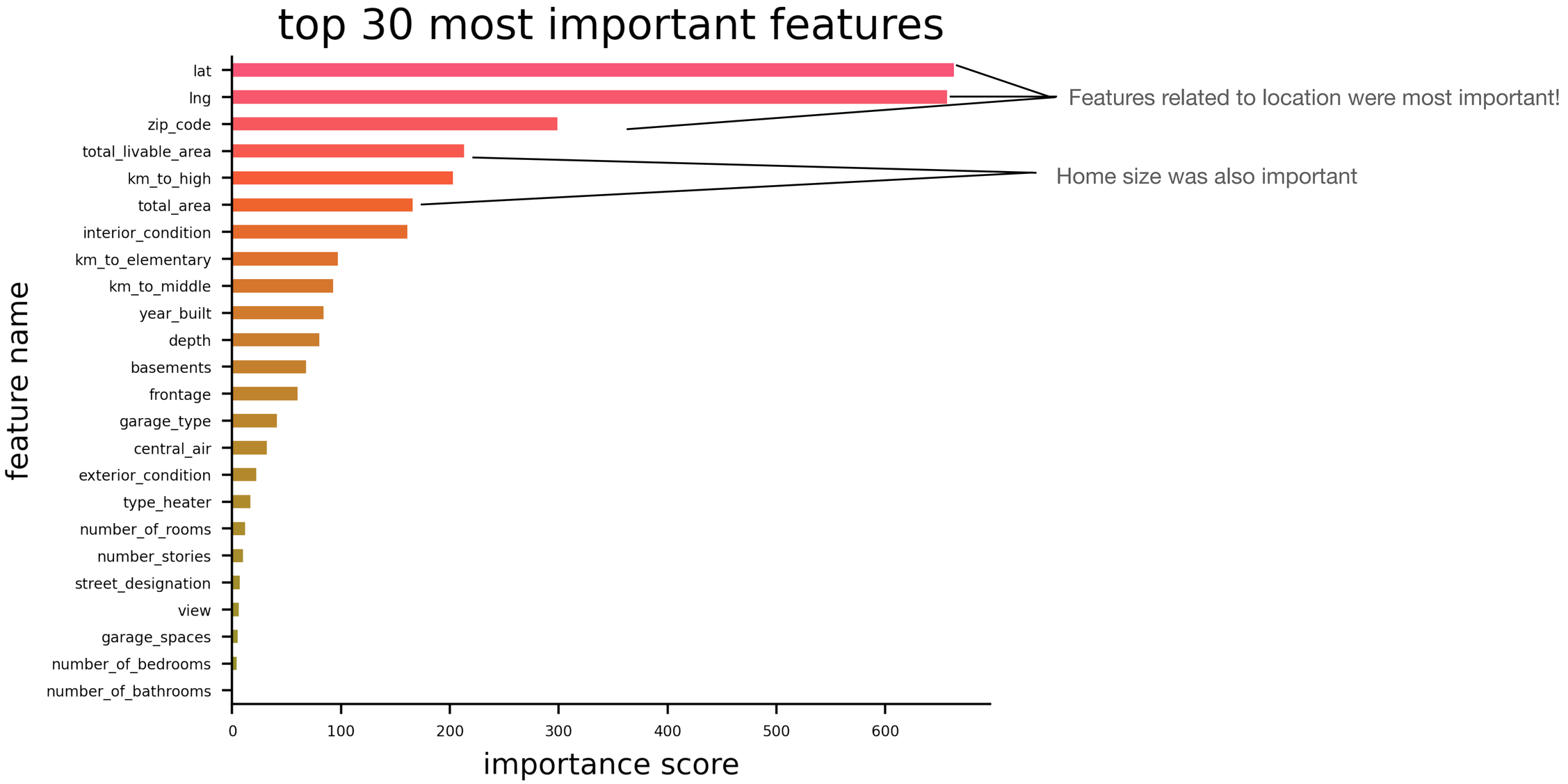
4. Computed MSE

Repeated above 100 times

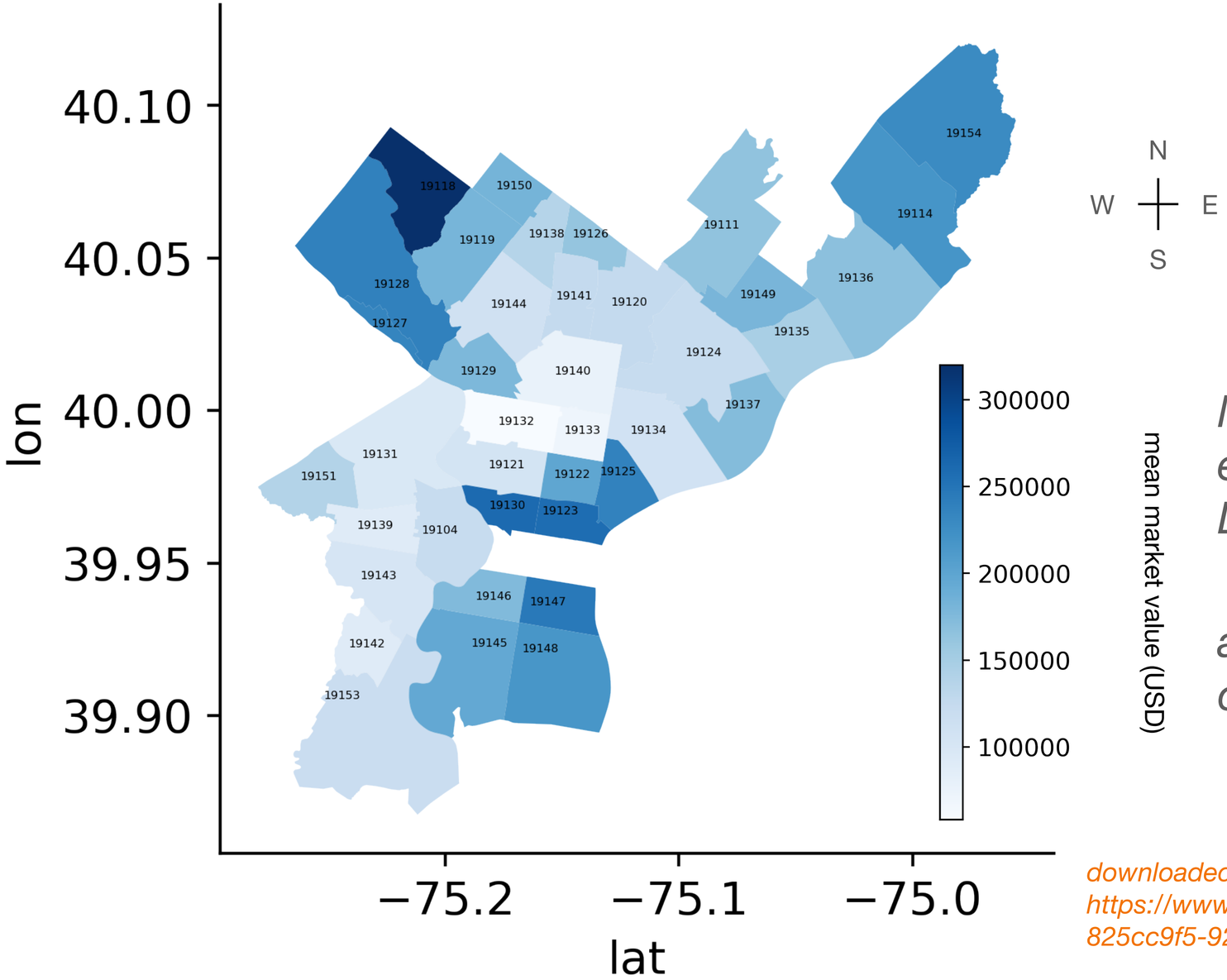


MSE for best model was lower than all 100 shuffled models, suggesting chosen model has predictive power

Which features were most important for predicting market value?



Market value by zip code in Philadelphia county

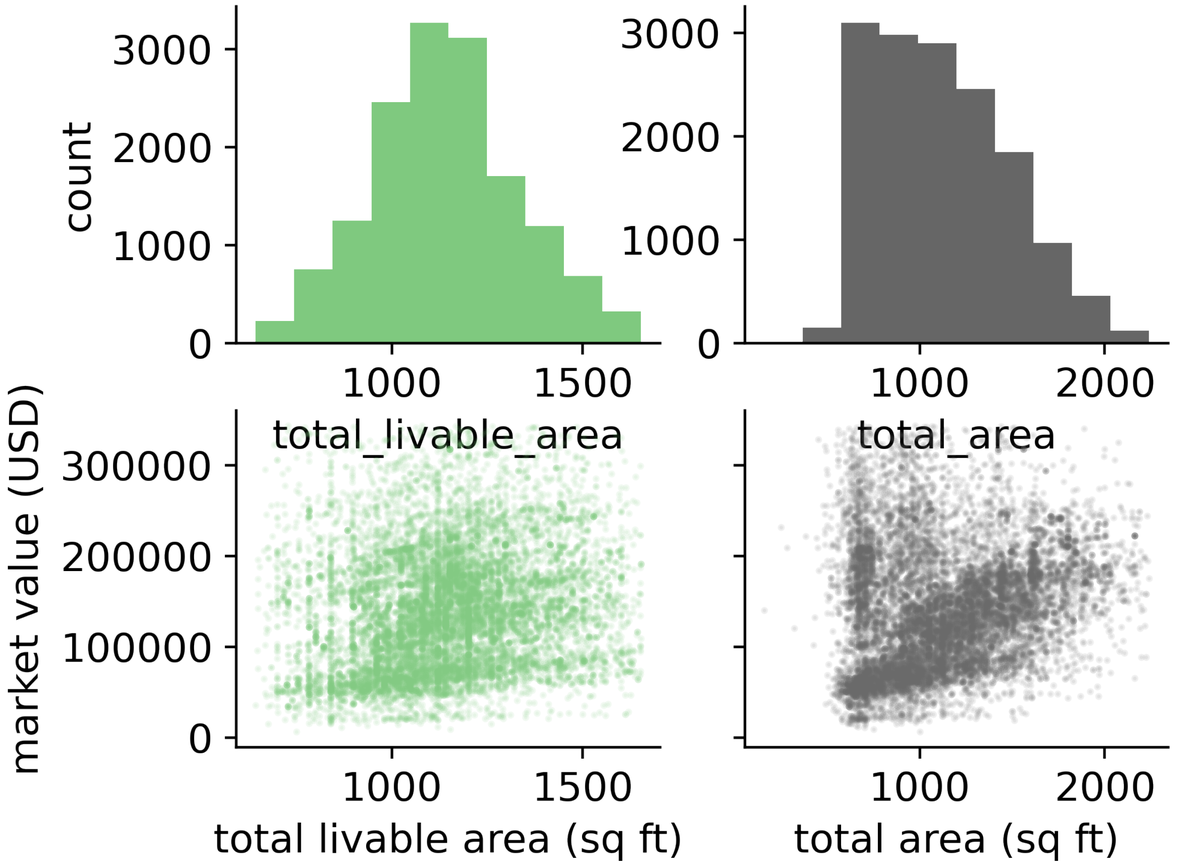


Market value is generally higher on the east side of the city, closer to the Delaware river

and lower in the west and the north central regions

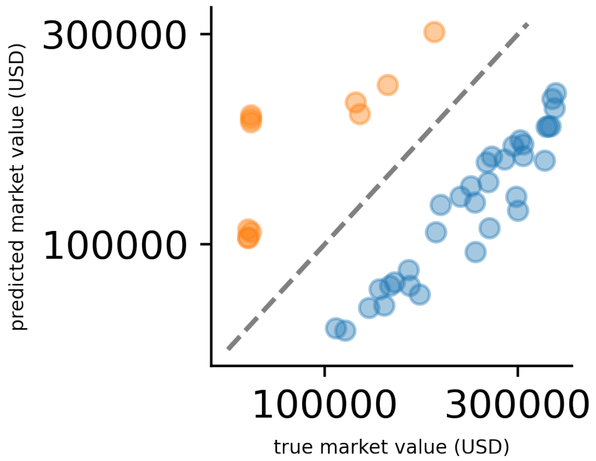
downloaded geoson data to produce map from OpenDataPhilly https://www.opendataphilly.org/dataset/zip-codes/resource/825cc9f5-92c2-4b7c-8b4e-6affa41396ee

Another look at square footage and its relationship with market value



Market value is higher for higher square footage in general.
There's a noticeable cluster of homes with higher property values but lower total square footage

Where does the model make the biggest errors?



top 99% of errors (45 worst predictions)

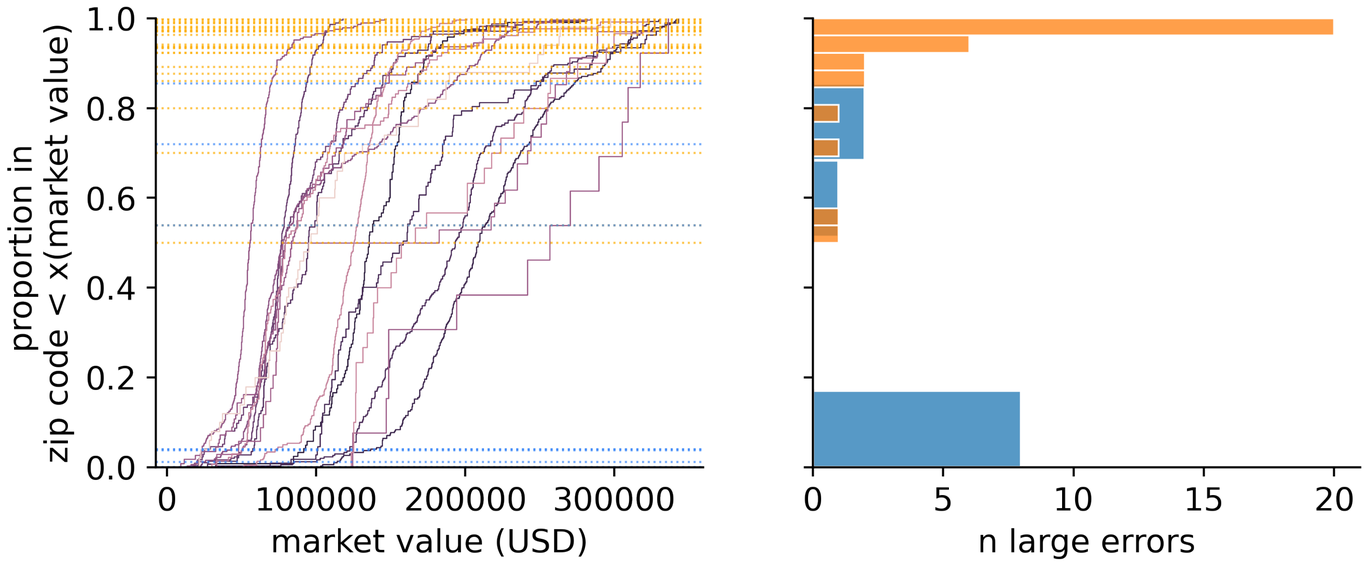
model overestimates market value

model underestimates market value

For the largest errors, the model underestimated market value more than it overestimated it

Is there something unique about the properties with largest error and where they are located (given that this was the biggest predictor of market value)?

Where does the model make the biggest errors?



Cumulative distribution of market value grouped by zip codes

- percentile within zip code of actual market value for underestimations
- percentile within zip code of actual market value for overestimations

Most large errors fell at the extremes in the market value distribution for its zip code

The model's prediction was closer to the middle of the distribution

What matters most when it comes to market value?

Location, location, location

Bigger homes tend to be more valuable
—expanding livable square footage could increase home value

Improvements to the interior condition of the home may also add to home value