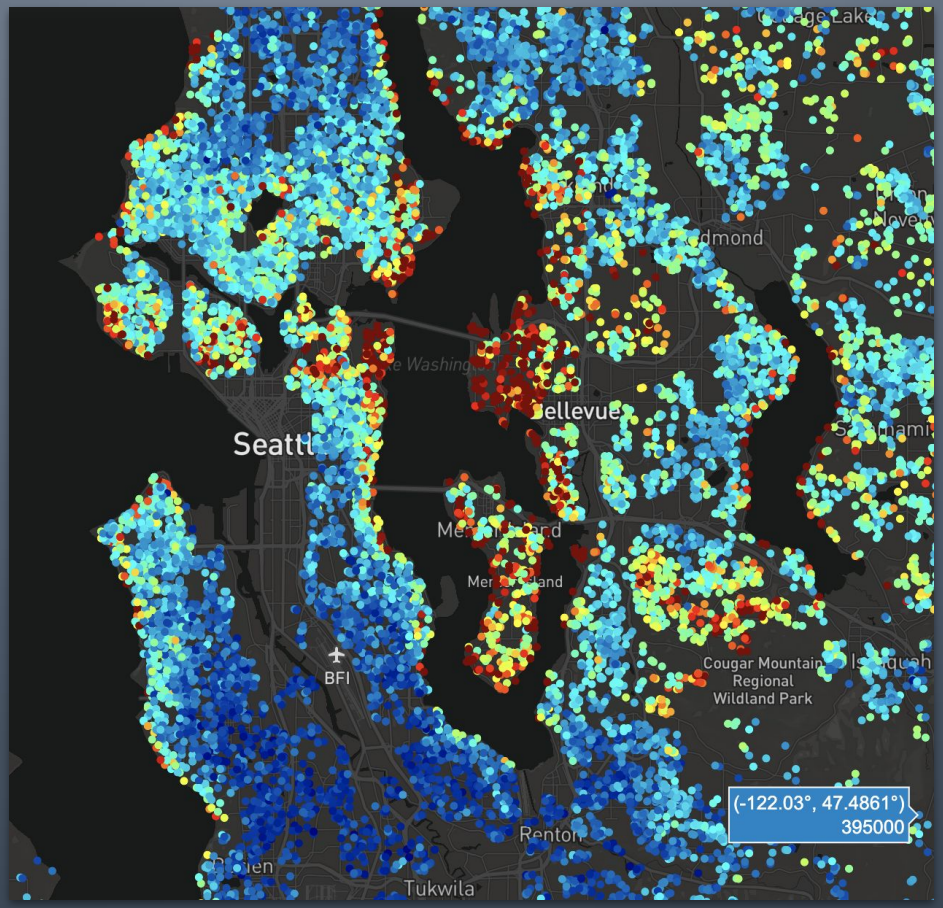
The Data Science of Selling your Home

Data-proven variables for maximizing profits

Background

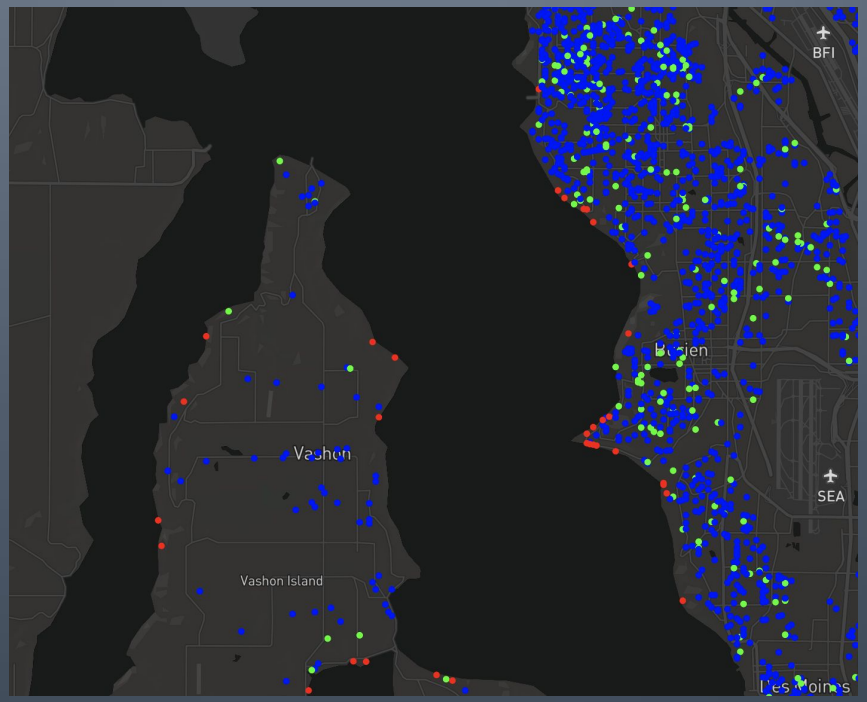
- Find variables contributing to a higher selling profit.
- Data from houses sold between 2014-2015 in King County, Washington State, USA.
- Based on 21,596 data points with 15 variables each.



Map made with Mapbox and Plotly. Showing price distribution in King County, Washington

Data Cleaning

- Kept all variables, only dropped ID.
- Converted date to month, day, and year columns.
- Filled NaN's with the mode of column.
- Dropped incorrect values (i.e. 33 bedroom house with 1 bathroom)
- Categorized
  - waterfront, view, grade, condition, zip
- Scaled
  - sqft_living , sqft_above, sqft_lot, sqft_living15, sqft_lot15, sqft_basement (boolean transform)
  - yr_built
  - yr_renovated (boolean transform)
  - price (log scaled)



Map made with Mapbox and Plotly. Showing incorrectly labeled waterfront properties in green as NaN's.

Methodology

Clean Data

Train_Test_Split

Run 4 Models

grid search for ideal model parameters

R^2, adjusted R^2, mean squared errors

Choose one model for easier interpretation of coefficients

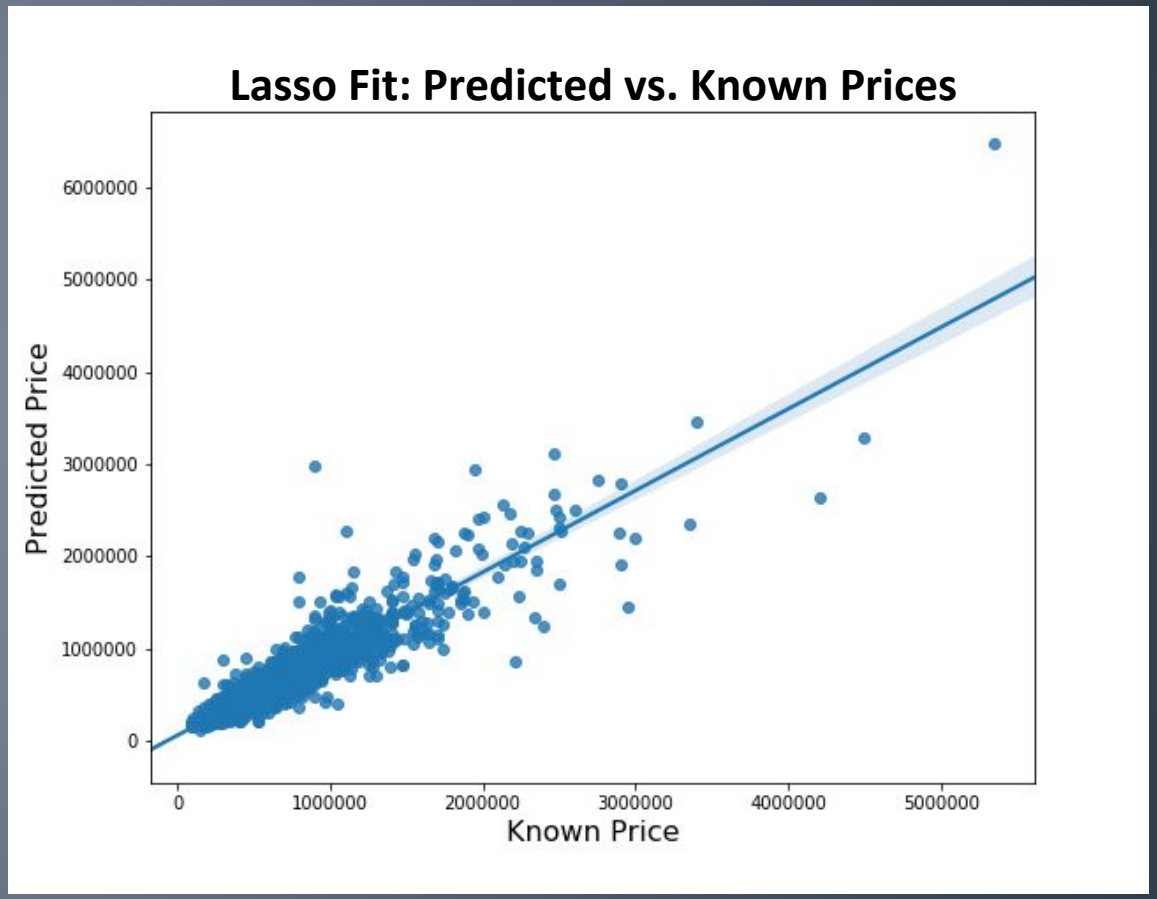
cross-validation with K-folds and obtained p-values to verify robustness of model

Predicted on the test data

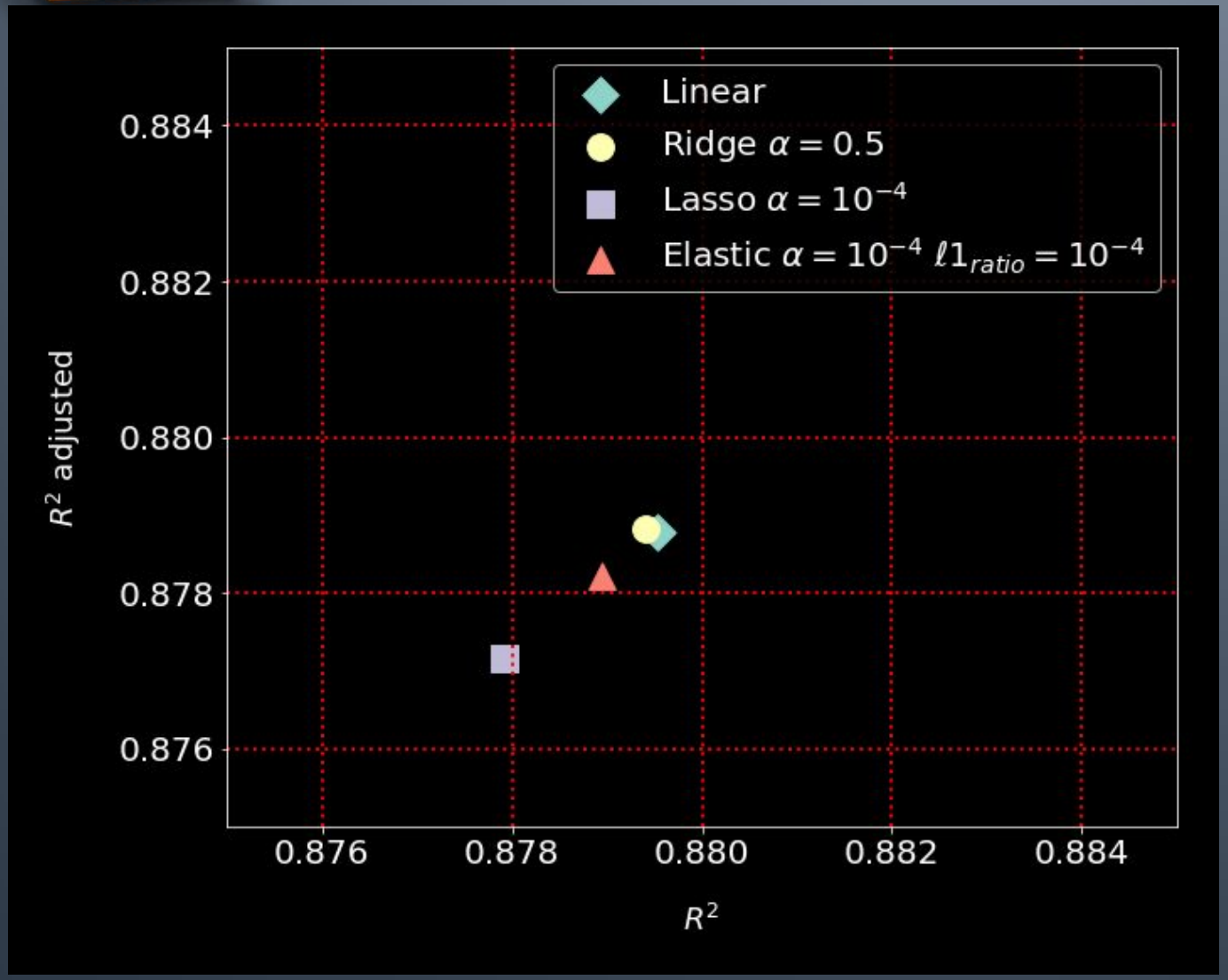
Visualized and interpreted results

Linear Regression
Ridge Regression
Lasso Regression
Elastic Regression

Models Used



Model Comparison



Model predictions for higher price

- Higher latitude
- Living by the water
- Increasing square footage
- Having a bigger yard
- Living in the 98039 zip code (Medina, Washington)
- Having rich neighbors like Bill Gates

Increasing Grade

Grade Scale is from 1-13 and can be found here.

Initial Grade	Final Grade	% Increase in price
4	7	16%
4	10	51%
10	11	1.5%

Adding View

Initial View	Final View	% increase in price
0	1	13%
0	4	32%
1	3	8%
1	4	18%

Adding Square Footage

Future Work

- Remove expensive houses from data (outliers).
- Try binning some of the categorical data.
- Separate regression for different subsets.

Questions?