Austin Housing Price Regression

Oscar Ko

Contents

The Data

- 15.2 thousand house records from Zillow home listings
- 47 features including:
    - Tax rates, garage spaces, cooling, heating
    - appliances, lot size, living area size
    - number of schools, bathrooms, bedrooms, and stories
- Dataset found on Kaggle
    - https://www.kaggle.com/datasets/ericpierce/austinhousingprices

The Goal

- Creating regression models
  - Selecting the best one for predicting housing prices in Austin, Texas

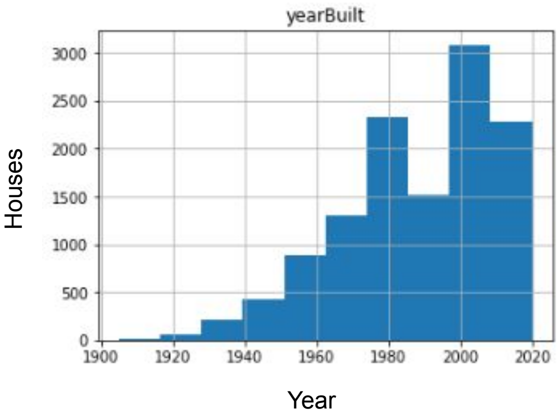
Conclusions: Best Model

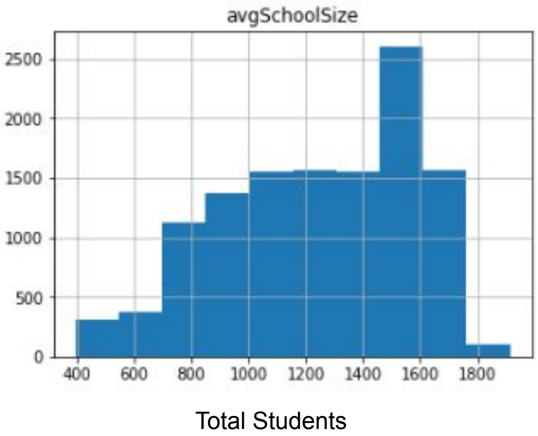
- Linear Regression
  - 31 Features that remained after Backwards Elimination
  - Train set R-squared = 0.46
  - Test set R-squared = 0.43

Conclusions: Other Models

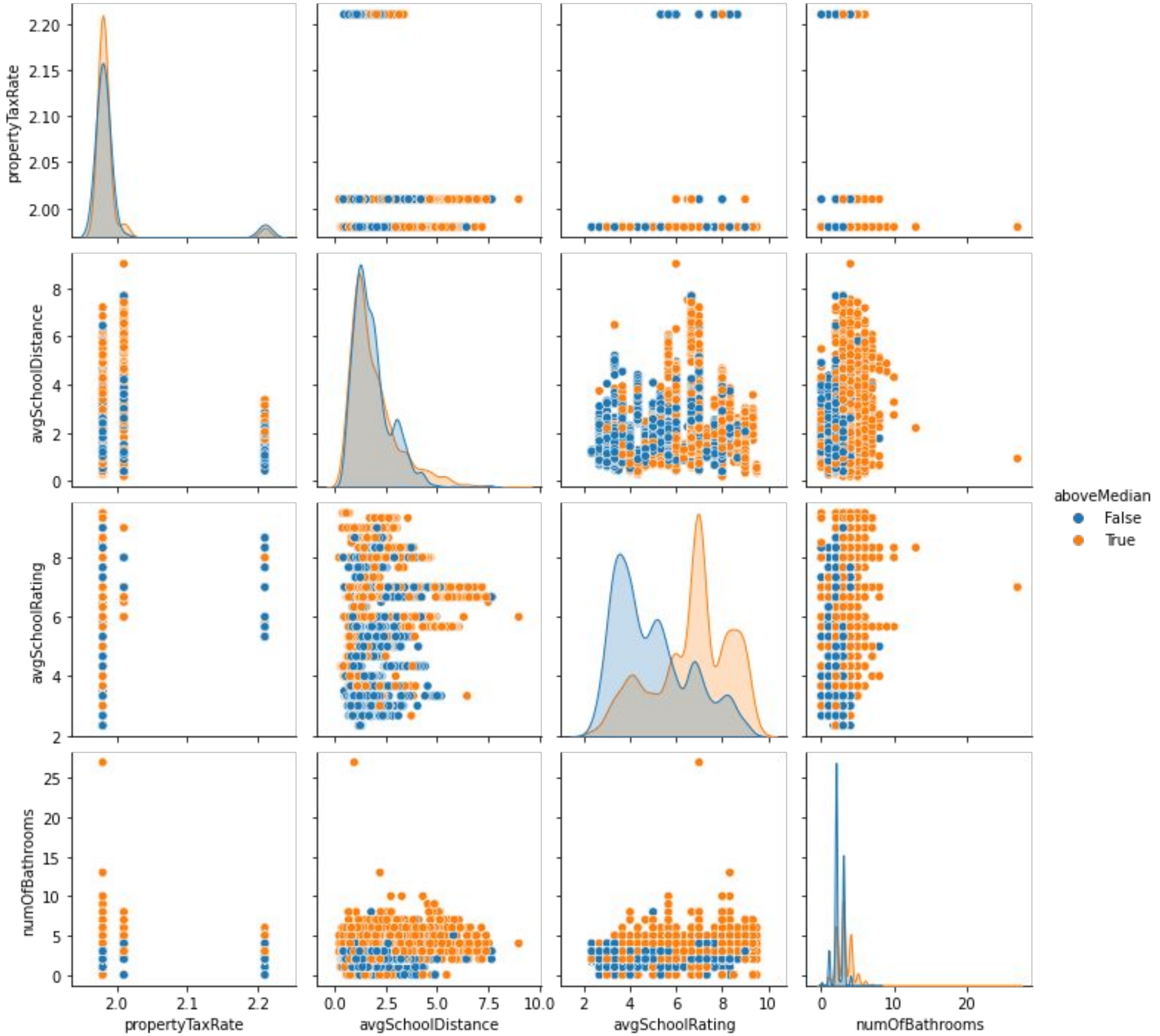
- What improved prediction
  - **Backwards elimination**
    - Reduced the number of features, and the R-squared remained the same.
- What didn’t improve prediction
  - **Dimensionality Reduction** (PCA)
    - The dimensionality reduction worked, but the model’s predictive power decreased:
      - (Without PCA) Train set R-squared = 0.46
      - (With PCA) Train set R-squared = 0.39
  - **Ridge** optimized with GridSearchCV
    - Train set R-squared = 0.41
  - **Lasso** optimized with GridSearchCV
    - Train set R-squared = 0.24

Other Findings: YearBuilt, SchoolSize, Bedrooms





Other Findings:
SchoolRatings & Bathrooms



Other Findings: Two Best Predictors

- Lasso
    - Just two features helped explain about 24% of the variability in the test set's housing prices
        - Living Area Square Feet
        - Number of Bathrooms

Other Findings: Two Best Predictors

- The best Linear Regression model (31 Features)
  - livingAreaSqFt coefficient = $106,300
  - numOfBathrooms coefficient = $224,300
  - For every 1 unit increase for each feature, the house's price will increase by the coefficient.
    - The features are standardized, a shift of 1 standard deviation to the right of the distribution.