Lotwize

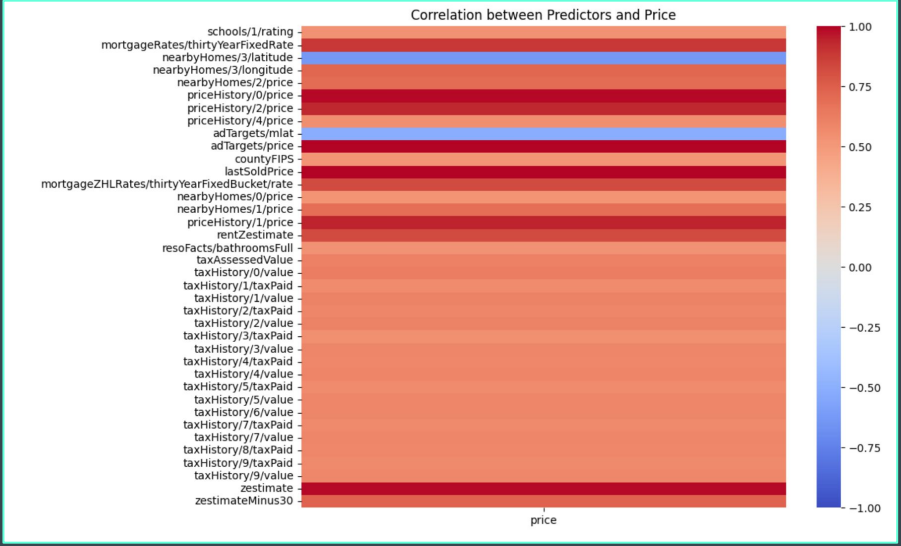
•••

Ryan Jewik

Problem Statement

Today the price of a property has never been higher, making the process of researching and buying a home the scariest it's ever been. Lotwize wants to make that process a bit easier by providing a tool that can reliably evaluate the value of a property based on statistical home data. This allows the potential buyer to make informed decisions and have confidence that they are getting what they pay for.

Preprocessing



- Initially too many variables (356)
- Took variables with correlation above 0.5 or under -0.5 to price
- Also manually reviewed columns and explored
  - Data types
  - Unique values
  - Counts
- Added additional variables such as:
  - Bathrooms
  - State
  - Square footage
  - Home type

Feature Engineering and Selection

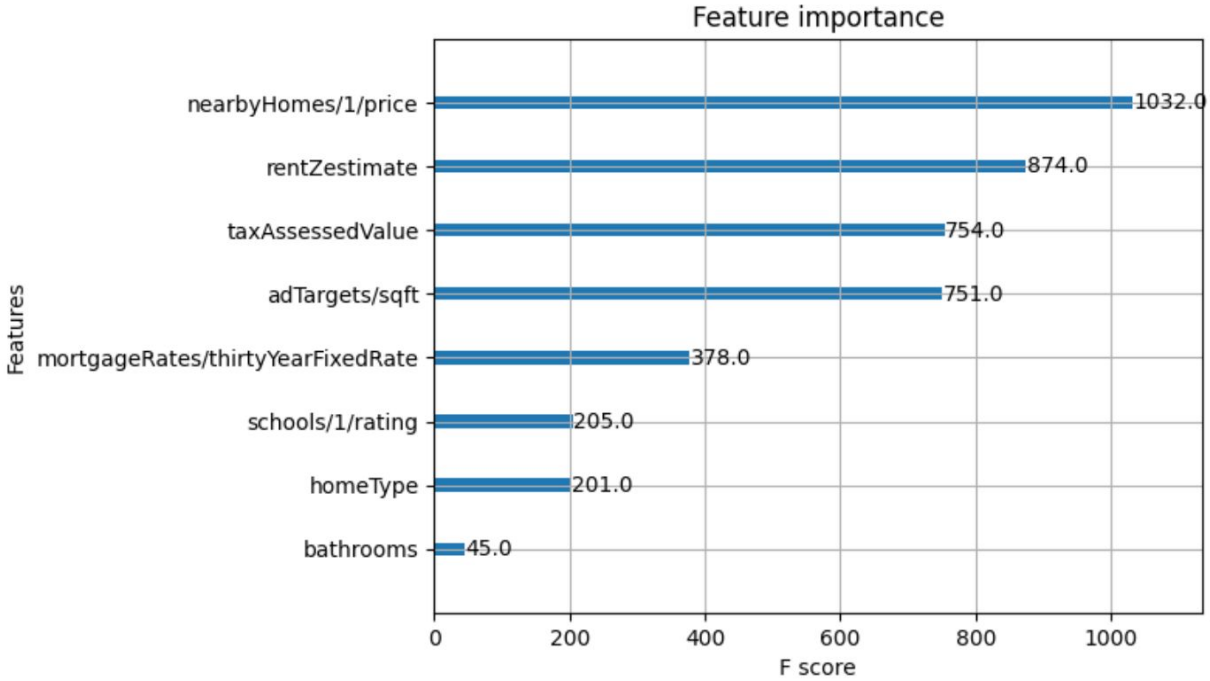
<bound method NDFrame.describe of schools/1/rating mortgageRates/thirtyYearFixedRate nearbyHomes/1/price
0 9.0 6.194 2700000.0
1 9.0 6.194 1870600.0
2 8.0 6.194 2295500.0
3 9.0 6.194 1250000.0
4 7.0 6.194 1565100.0
.. ..
766 6.0 5.879 664400.0
767 7.0 5.879 599900.0
768 7.0 5.879 744000.0
769 6.0 5.879 674400.0
770 7.0 5.879 650000.0

rentZestimate resoFacts/bathroomsFull taxAssessedValue bathrooms \
0 6088.0 2.0 1400439.0 2.0
1 5872.0 3.0 1021308.0 3.0
2 6999.0 3.0 953595.0 3.0
3 7016.0 2.0 321945.0 2.0
4 7701.0 3.0 1404540.0 4.0
.. ..
766 2792.0 2.0 348167.0 2.0
767 3280.0 2.0 535806.0 2.0
768 3376.0 2.0 81705.0 3.0
769 2797.0 2.0 502859.0 2.0
770 3459.0 2.0 216719.0 2.0

homeType resoFacts/bathrooms state adTargets/sqft price
0 CONDO 2.0 CA 1493.0 1600000
1 SINGLE_FAMILY 3.0 CA 2163.0 2050000
2 SINGLE_FAMILY 3.0 CA 3472.0 2340000
3 CONDO 2.0 CA 1975.0 1625000
4 SINGLE_FAMILY 4.0 CA 4283.0 1642000
.. ..
766 CONDO 2.0 CA 948.0 645000
767 CONDO 2.0 CA 1225.0 600000
768 TOWNHOUSE 3.0 CA 1240.0 615000
769 CONDO 2.0 CA 948.0 640000
770 TOWNHOUSE 2.0 CA 1182.0 635000

- The bathroom variables were the same so we removed most of them
- State was entirely California, no other states included in the data set
- Other variables like price history were collinear with price
- Ratings could have potentially been categorical however there's a chance two schools could have the same rating

Problem Solving and Analytics



- Other variables regarding price such as Rent Zestimate, nearby home prices, and price history were all highly important
- Bathrooms and bedrooms proved to not be (relatively) important, likely because a large number of homes with wide price variances only have 2-3 bathrooms
- Reduced features to 8 in final model

Predictive Modeling

Gradient Boosting Tree

- K-fold cross validation
- Due to overfitting reduced features
- Used early stopping to find optimal hyperparameters
- Still struggled with overfitting, likely due to features / data

```
Accuracy: 58.06% (36.53%)
Train MSE : 14587371408.55011
Test MSE  : 32638916396.510056
Train MAE : 89462.4019256795
Test MAE  : 137734.24157714844
Train MAPE : 0.0746505640614695
Test MAPE  : 0.14191590655567635
Train R2 : 0.9567492827773094
Test R2  : 0.5805641263723373
```

Linear/Polynomial Regression

- Additional polynomials worsen the results
- Normal train test split showed good R2 but error values metrics were exceedingly poor
- K-fold cross validation confirmed that

Train Test Split

```
Train MSE :  24623801046.136616
Train MAE :  102698.32367983983
Train MAPE:  24623801046.136616
Train R2  :  0.9376718451595064
Test MSE  :  30999133687.501564
Test MAE  :  99133.73653339202
Test MAPE :  30999133687.501564
Test R2   :  0.920331081124101
```

K-Fold Cross Validation

```
Train MSE : 25043742275.727337
Train MAE :  102798.82808332695
Train MAPE:  26042614908.55837
Test MSE  : 30817603129.67369
Test MAE  :  129975.6674359958
Test MAPE :  25005749465.119358
Train R2 : 0.9354114118907304
Test R2  : -0.21022450336058082
```

Business Impact

In the interest of creating the best possible Automated Valuation Model, features centering around price such as nearby housing prices, tax assessed value, and mortgage rates have the highest importance in the model. Features that don't very often and cannot be clustered well due to high variance such as bathroom count, bedroom count, and state don't bring as much importance to the model.

Appendix

- Used correlation graphs and filtered them to find important variables
- Did additional manual analysis to pick out more variables that I thought to be potentially important
- Initially began with polynomial regression, but soon found adding polynomials made the results significantly worse
- one hot encoded categorical variables and z-scored continuous variables
- Used MSE, MAE, MAPE, R2 as evaluation metrics
- In an attempt to resolve overfitting issues with both the linear regression model and the gradient boosting tree, I used K-fold cross validation
- Used Gradient Boosting Tree next but had similar (but not as drastic) overfitting issues
- Used early stopping to find optimal hyperparameters and reduce overfitting
- Used identical metrics to evaluate model
- Tuned parameters using SHAP values and Feature Importance but it oversimplified the model