HCAD House Value Prediction

By Rafael Pinto Ortiz

Harris County

Harris County is the third-most populous county in the USA.

Its appraisal district (HCAD) provides a fantastic dataset with each appraised property characteristics (appraised value, fixtures, features) reported yearly.

4 Million people live in Harris County

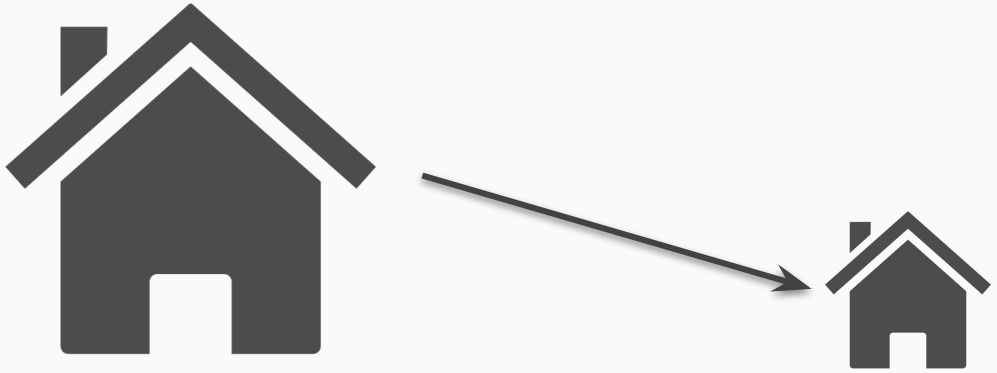
https://www.tshaonline.org/handbook/entries/harris-county

Value prediction

Using HCAD data, I'll follow the Data Science Method to understand if my house was fairly appraised in 2016.

Then, I'll construct a statistical model for predicting house appraised values based on the house features (area, number of rooms, pool, etc.).

Was my house fairly appraised in 2016?
If not, what is a fair value?



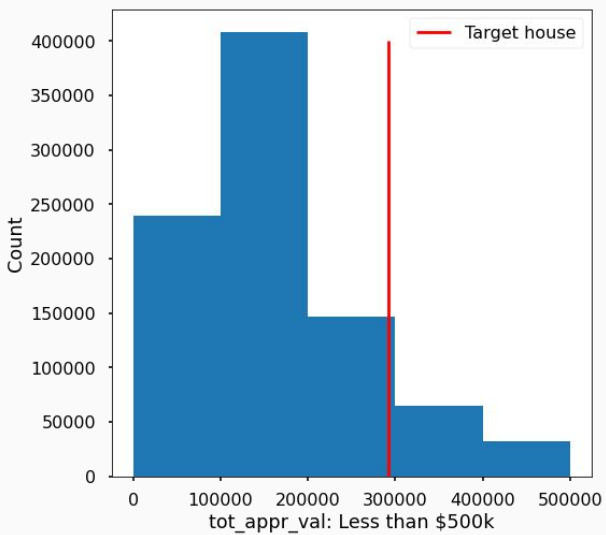
HCAD
$292,707

My model
$258,696

All Houses

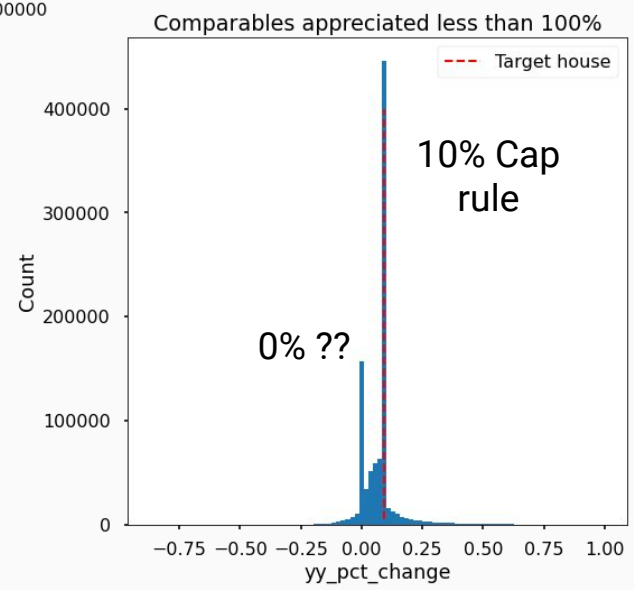
There is a wide range of property values given their location (neighborhood), physical condition, renovation, and other factors.

Let's calculate the year-to-year percent change in property value, in an attempt to minimize the effect of the property's appraisal value magnitude.



A wide distribution
1 Million properties

Interesting spikes



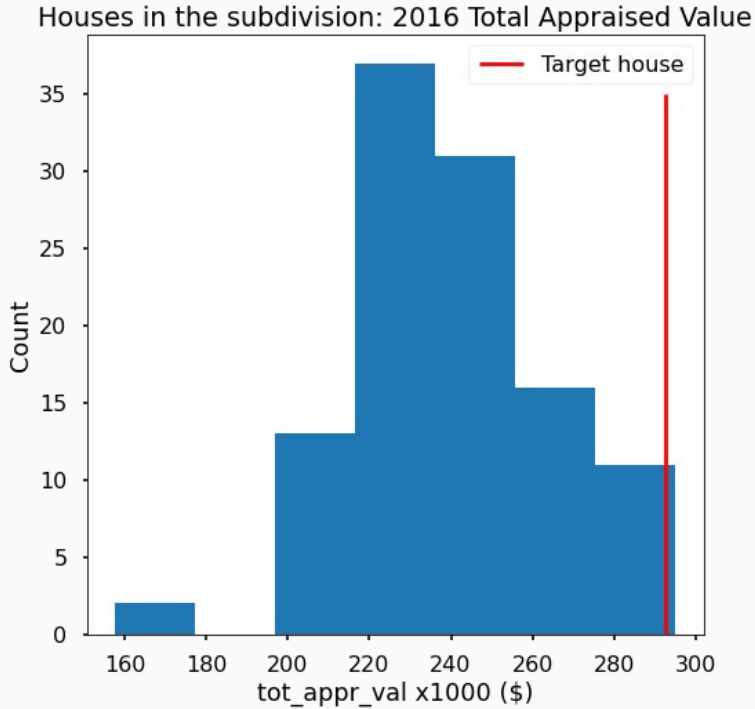
Subdivision Homes

I selected similar properties to my target house by counting only properties in the same subdivision to narrow down the variability in the home values.

The target house sits on the high end of the distribution of appraised values for the houses in the subdivision.

Not all the subdivision houses are the same. To make a fair comparison, we should find the properties with similar characteristics to the target property,

110 properties
32 features

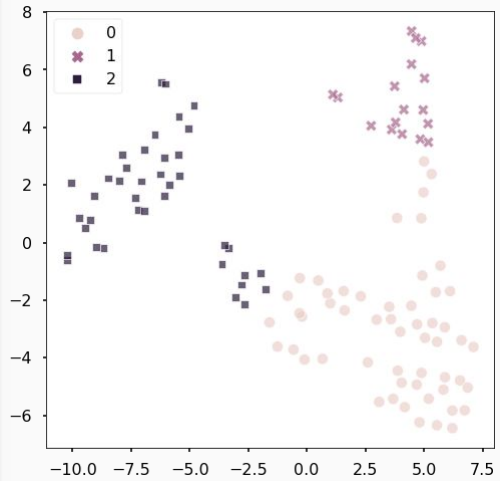


Comparables

I use k-means to group the subdivision homes based on their physical characteristics (area, rooms…)

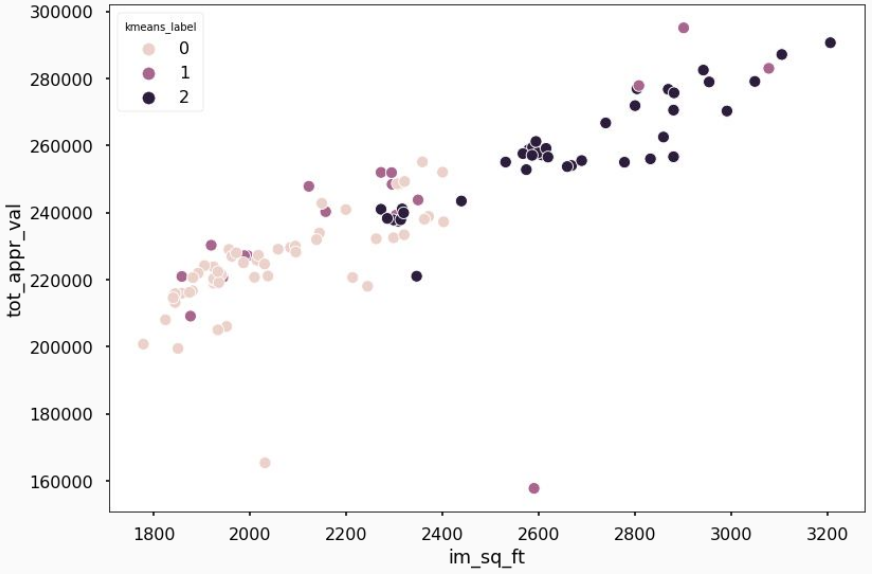
The k-means groups 0 and 2 represent houses with large area and value, and houses with low area and value, respectively

Also this figure shows the strong positive correlation between home area and appraised value



TSNE of K-Mean labels

Value vs Area



Null-Hypothesis

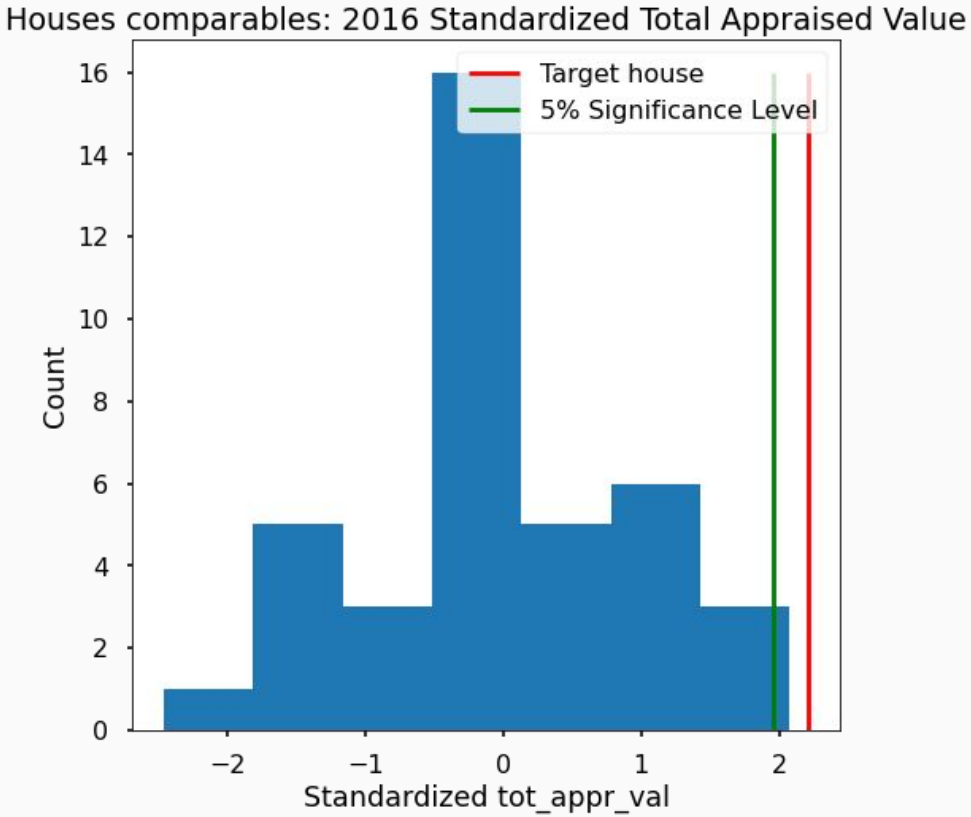
Population: Comparable houses (group 2) appraised by HCAD in the target subdivision.

Significance level: 5%

H_0: The 2016 appraised value for the target house was fair relative to its comparables.

H_a: The 2016 appraised value was unfair, or it doesn't belong to the comparables distribution.

I reject the null. What is a more appropriate appraised value?



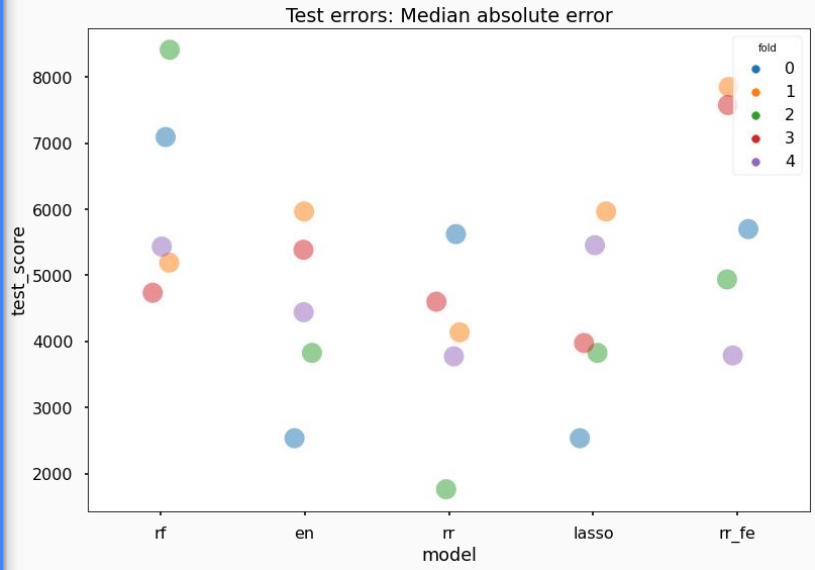
z-score: 2.21
p-value = 0.014

5-fold CV

Train set: 70%

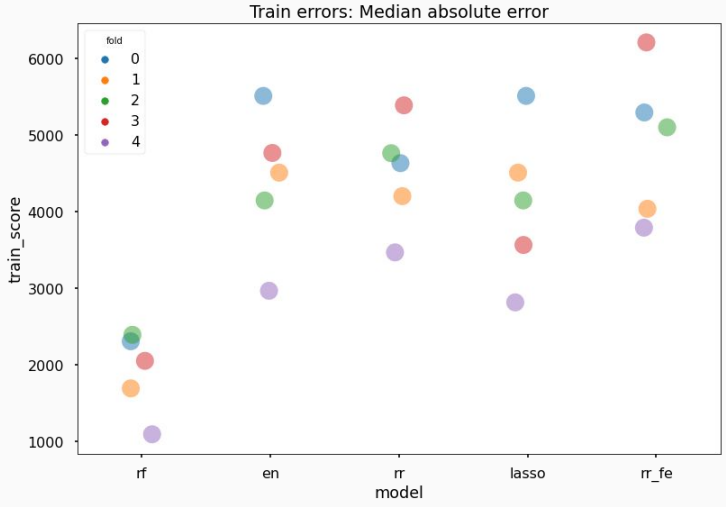
1. Fit a random forest (rf)
2. Standardize the variables
3. Fit an ElasticNet (en)
4. Fit a ridge regression (rr)
5. Fit a LASSO for feature elimination (lasso)
6. Fit a ridge regression with features passed by LASSO (rr_fe)

Nested CV for hyperparameter tuning.



Selection:
Ridge Regression has the smallest minimum and maximum cv-error

Random Forest is overfit

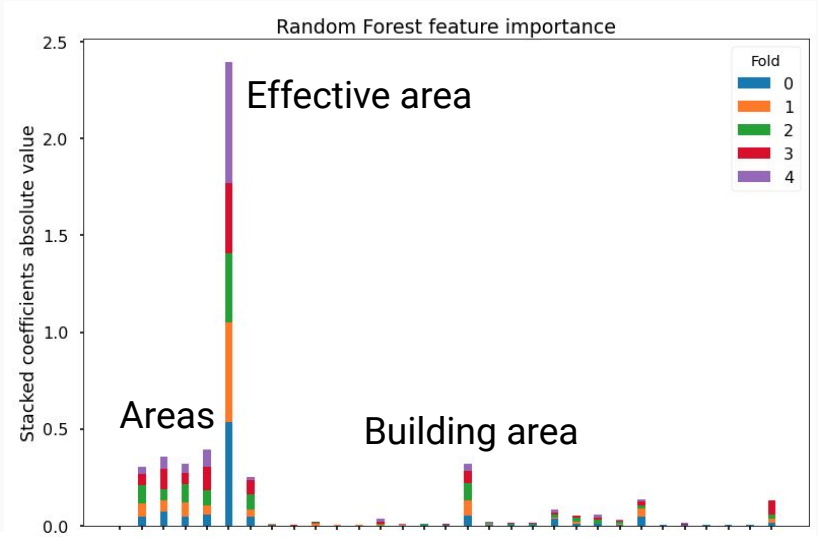


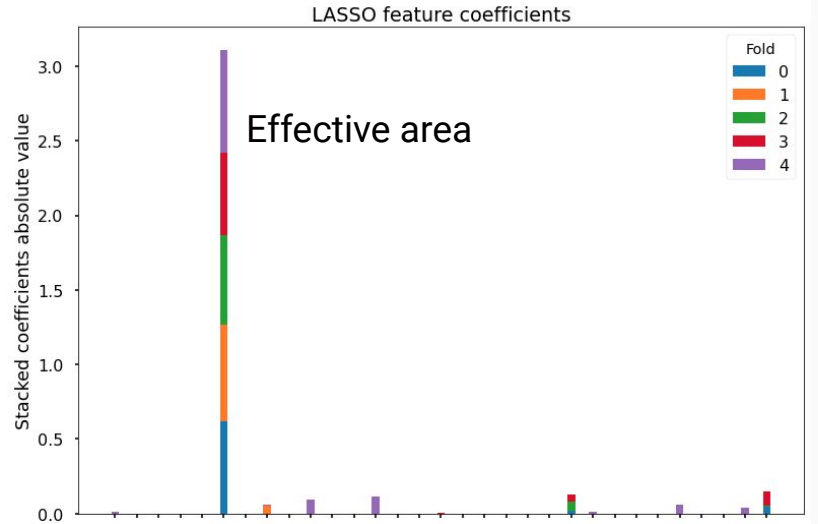
FE comparison

The ridge regression estimator outperformed the ridge regression with feature elimination.

But I wanted to understand the features importance.

Random forest and LASSO found that Effective Area is by far the most important feature in determining the property's value

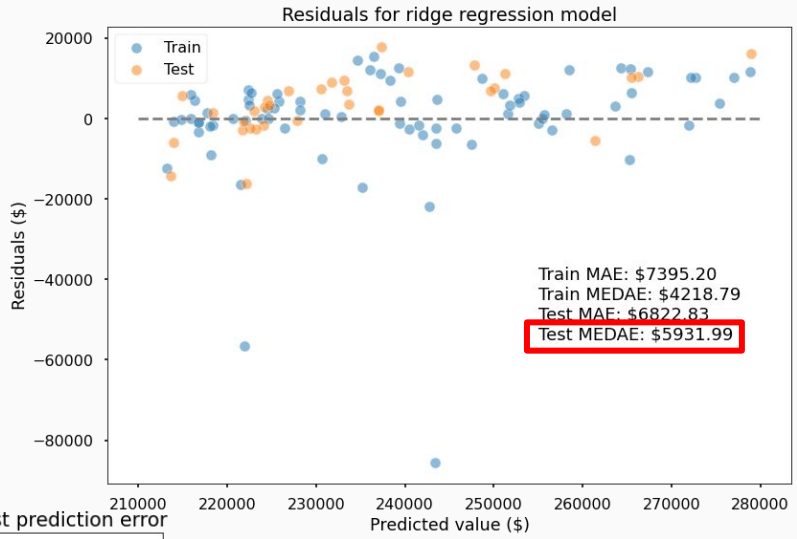


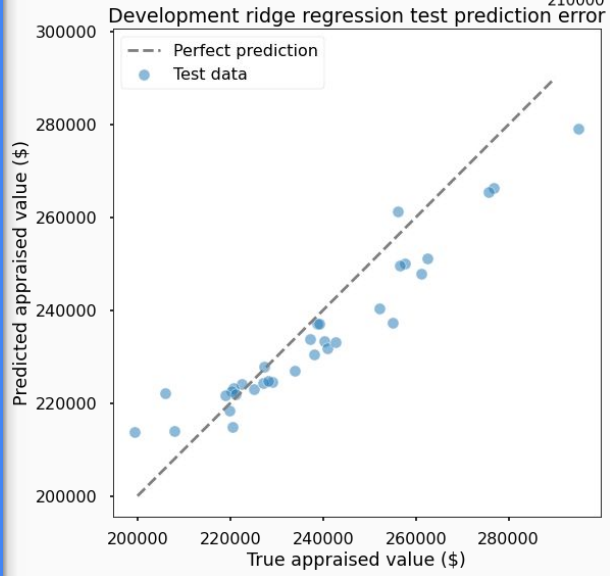


RR Errors

Test data: most residuals are positive, so model tends to underpredict.

Expected error: USD 5,932 on average according to the test set median absolute error (MedAE)





$5008

Is the difference between my model prediction ($258,696) and HCAD value after protest ($263,704). This residual is below my expected MedAE ($5930)

Conclusions

I showed with a hypothesis test that my 2016 HCAD appraised home value was unfair (p-value 0.014).

I built a ridge regression model to estimate the appraised value based on the houses in the neighborhood.

My prediction was $34,011 less than HCAD's, and $5,008 less than HCAD's after protest.

The model's expected median absolute error is $5,930.

First check the area hasn’t changed

Even if there is a mismatch you would still need to provide an expected value ➡ this model

https://unsplash.com/photos/r3WAWU5Fi5Q

Thanks!

To HCAD for a fantastic dataset.

Tommy B. (Springboard) for insightful discussions.

Project repo: https://github.com/RafaelPinto/hcad_pred

Reach out to:

https://github.com/RafaelPinto/

https://www.linkedin.com/in/rafaelpintoor/

https://unsplash.com/photos/p-rN-n6Miag