Prediction of Single-Family Home Appraisal Value and Valuation Appeals

Capstone 2

Springboard Data Science Career Path

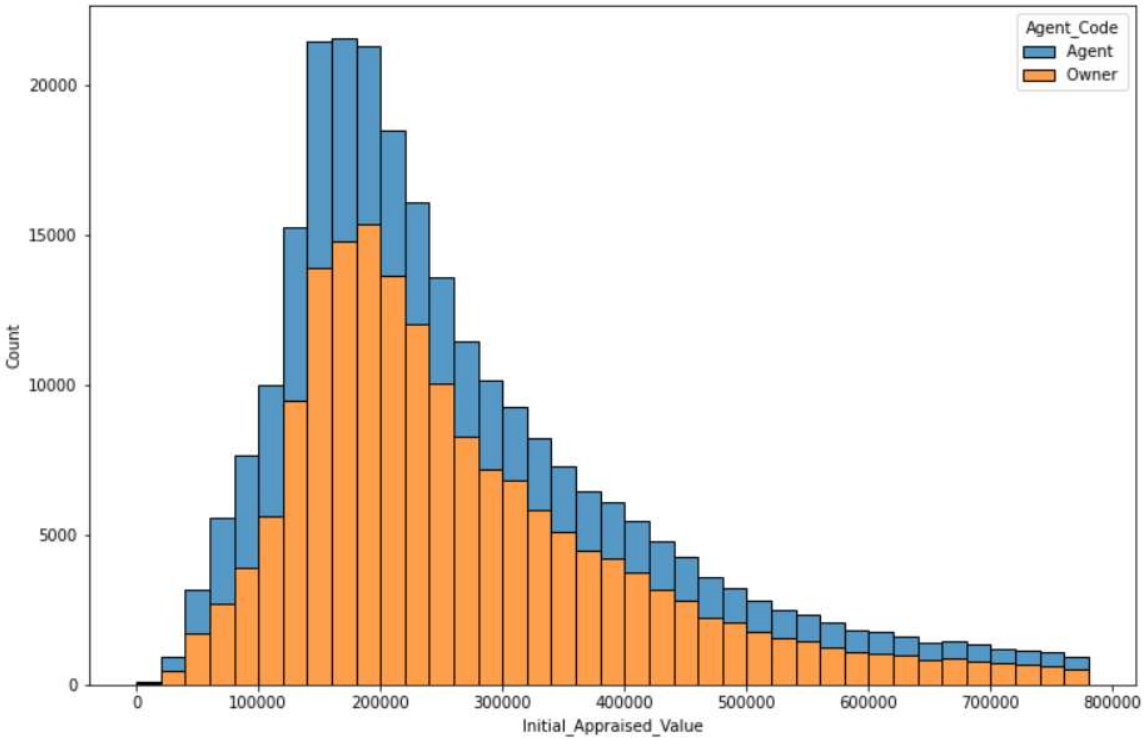
Kevin Wolfe

Harris County Appraisal District (HCAD)

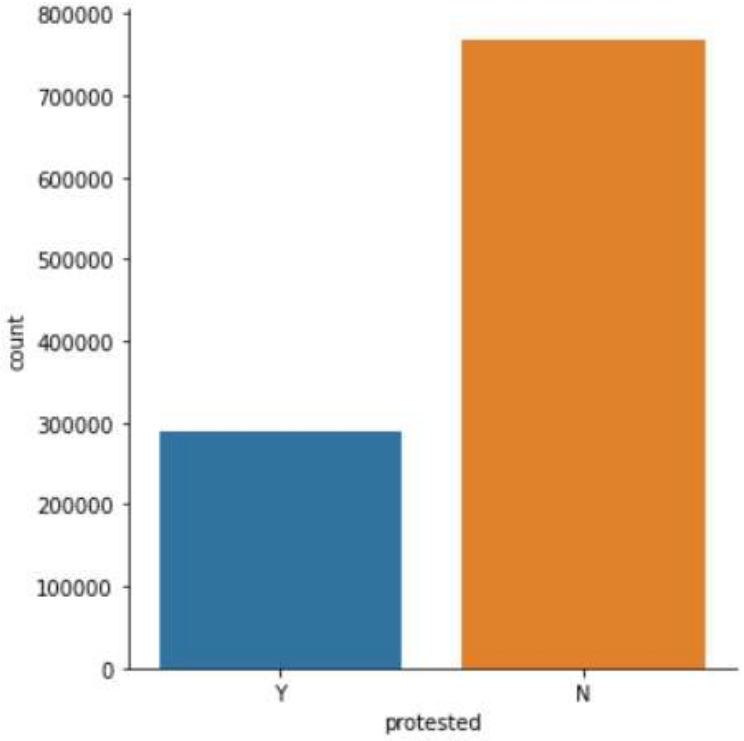
- Yearly appraisals are generated by HCAD for the 1.8 million property in Harris Co. TX
- Nearly 80% of single-family home properties experienced an increase in appraised value
- Property taxes are calculated based on this valuation
- The public can appeal their appraisal
  - 1/3 of properties appeal
  - 68% employ a 3rd party to manage this appeal

Using HCAD data can a model be created to predict appraisal value?

Using HCAD data can protests be predicted?



Protests Cost Money



Nearly 1/3 of single-family home appraisals are protested

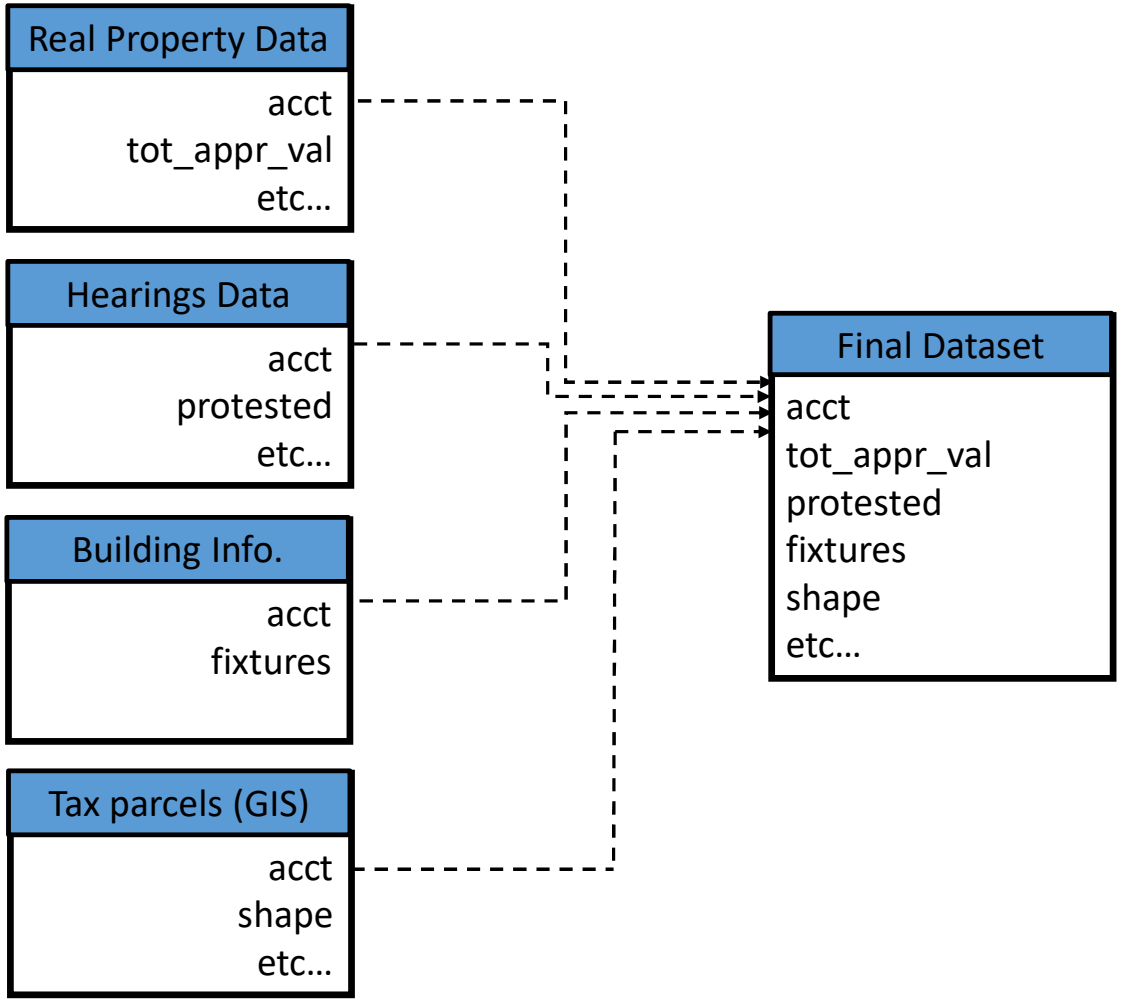
- Self-represented owners spend time on their appeal
  - Unsuccessful protests can be avoided by understanding appraisal.
- 3rd party managed appeals cost the owner
  - Flat rate
  - % of savings – can be hundreds of dollars
  - A better understanding of protests provides a business opportunity to 3rd parties
- HCAD holds over 400,000 hearings per year to resolve protests
  - Nearly 300,000 of these hearings are associated with single-family homes
  - Reducing these hearings would decrease administrative costs

Data Availability

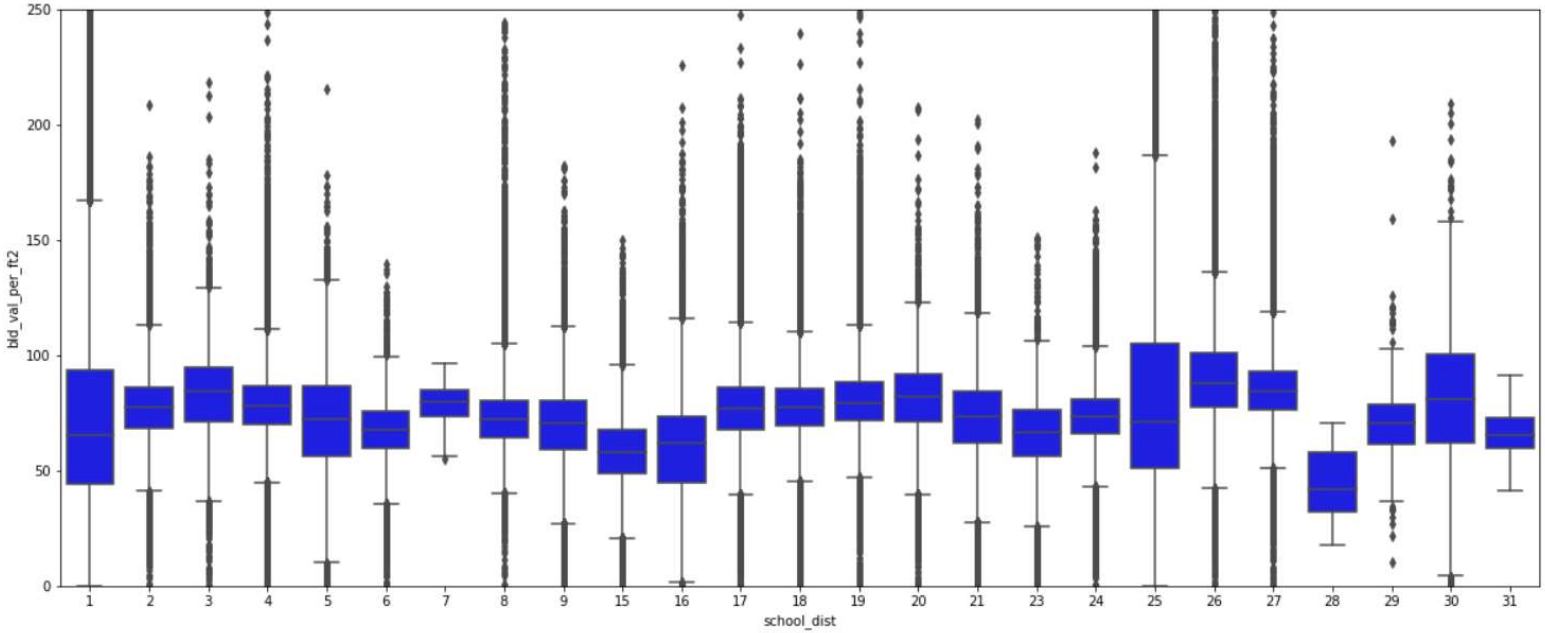
All the files used for this analysis and visualizations were 2020 datasets downloaded from https://download.hcad.org/data/

- Real Property Data - account and descriptive data about every property in Harris Co.
- Hearings Data - information about each of the properties that was appealed
- Building Information- information about the building
- Tax parcels - GIS data for each parcel in Harris Co. Utilized for map-based visualizations

These datasets were joined together by their common acct column prior to analysis.



Handling large categorical features



Categorical columns like school_dist were converted to a ratio factor so the large number of categories could be handled by machine learning algorithms.

$$\text{Factor} = \frac{Mean\ \$/ft_{catigory}}{Mean\ \$/ft_{all\ properties}}$$

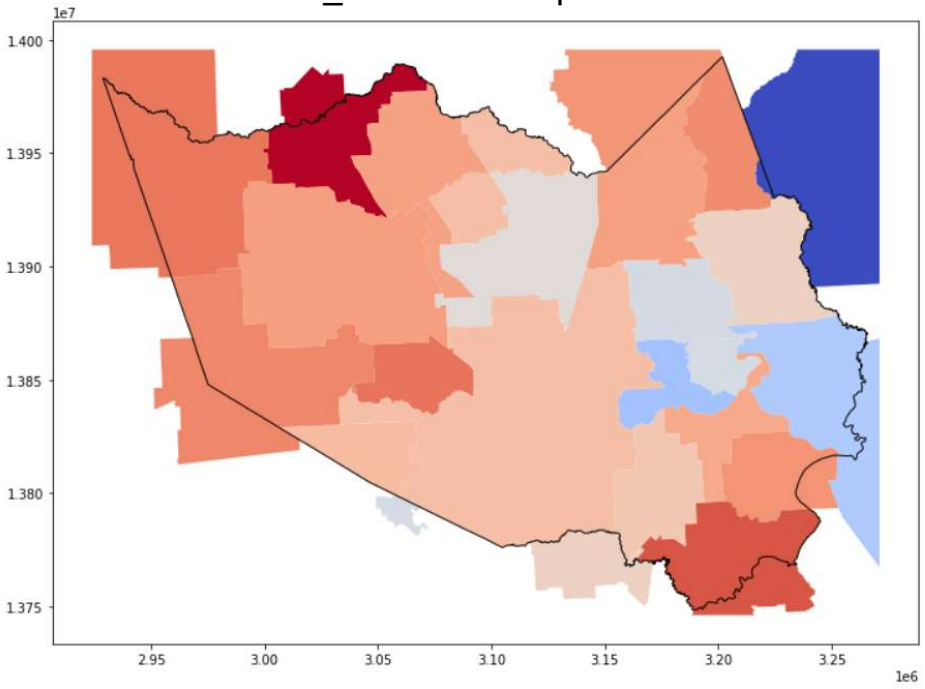
Category	School_dist	Neighborhood_Code	Neighborhood_Grp	Market_Area_1	Market_Area_2	Center_code
# Values	25	5900	926	162	141	32

Handling large categorical features (cont.)

Factors were created for both land and building for each of the large categorical features.

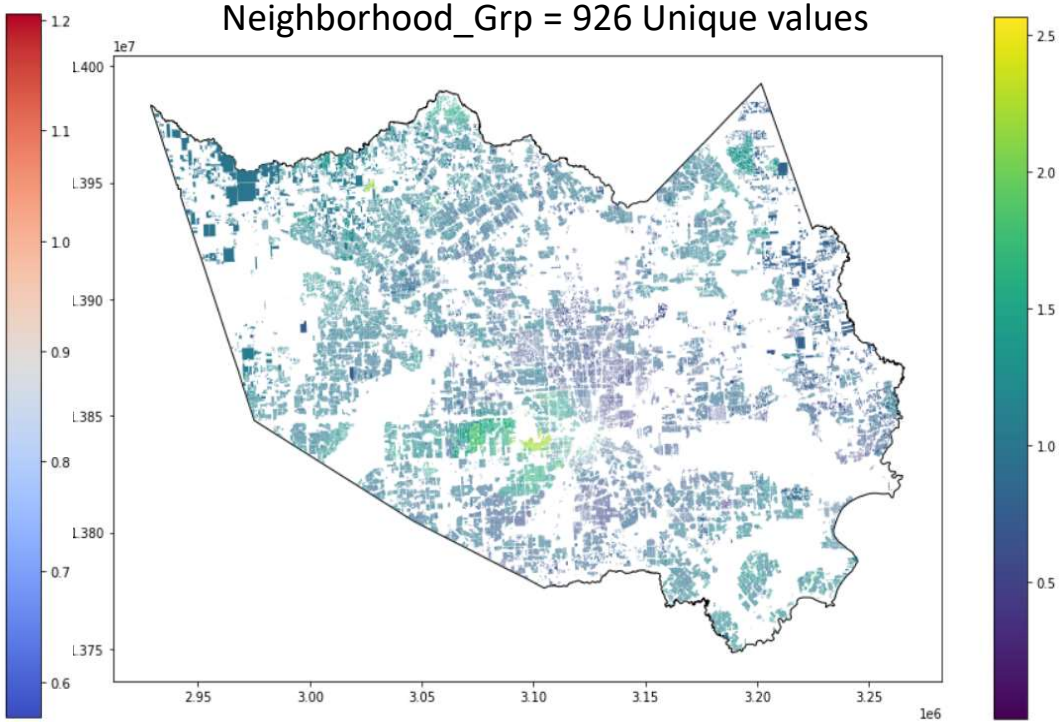
Each categorical factor had a different geographic range, in general the larger the number of categories the greater the range of values for the generated factor.

School_dist = 25 Unique values



School_dist_bld_factor

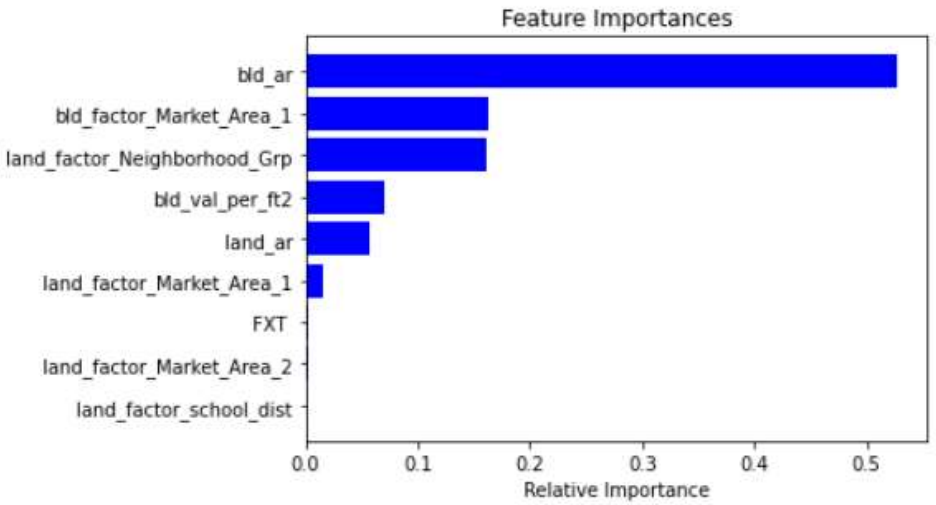
Neighborhood_Grp = 926 Unique values

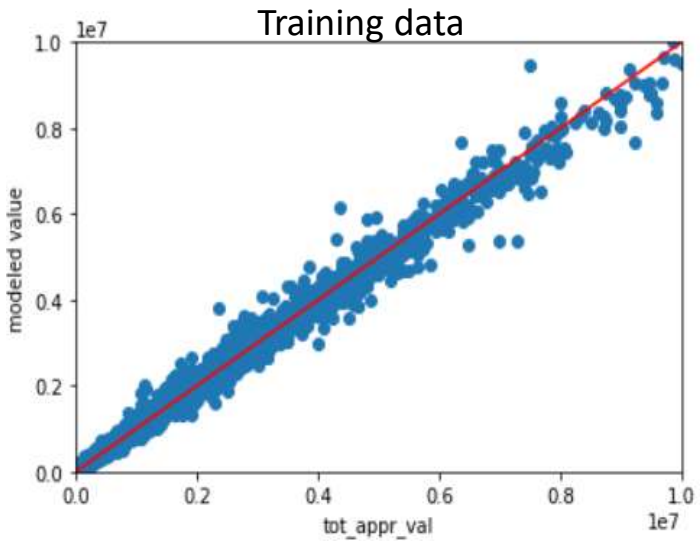


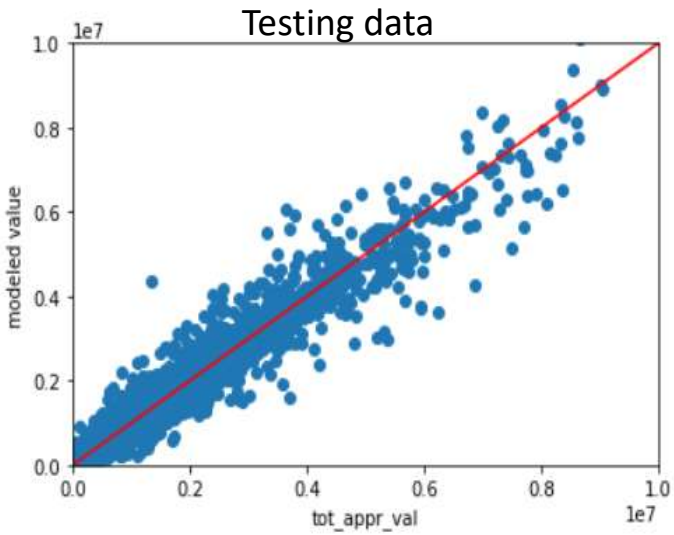
Neighborhood_Grp_bld_factor

Tot_appr_val modeling

- Using non-value based features can the total appraisal value be predicted?
- Linear models failed to handle more expensive properties
- The best model generated to date is a Random Forest Regressor
- Most impactful features were:
  - Building area (sqft)
  - Market Area 1 building factor
  - Neighborhood Group land factor
- R^2 Testing = 0.97
- RMSE = 50405







Protest prediction modeling

- Can a protest be predicted based on property data?
- The best model generated to date is a Random Forest Classifier
- Most impactful features were:
  - X_features_val – this was very low for most properties, but may be a trigger when it was > $0
  - Total appraisal value increase
  - Initial market value

Accuracy: 0.917

Balanced accuracy: 0.87

Precision score for "Yes" 0.916

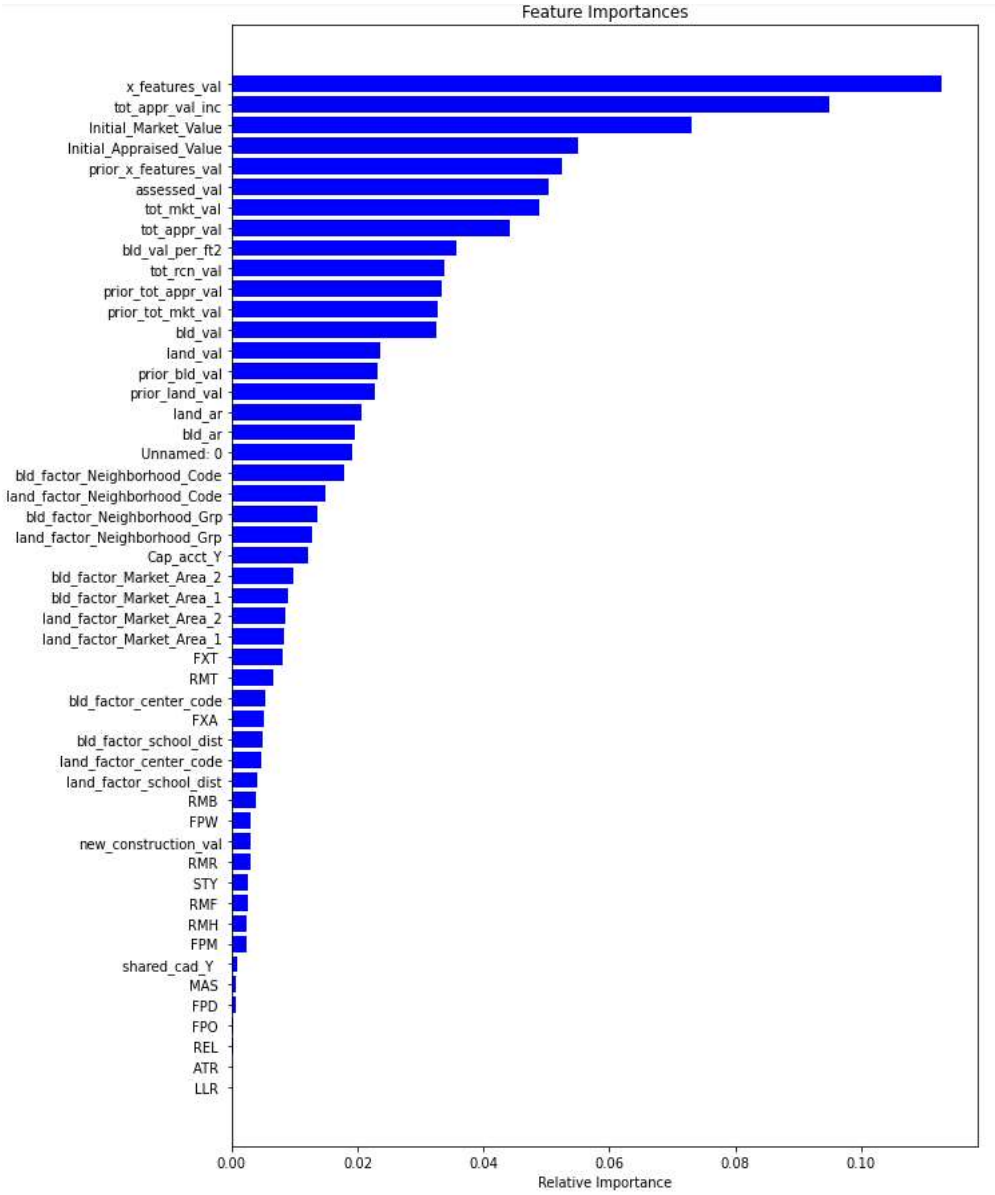
Precision score for "No" 0.917

Recall score for "Yes" 0.767

Recall score for "No" 0.973

Confusion Matrix – Test Data:

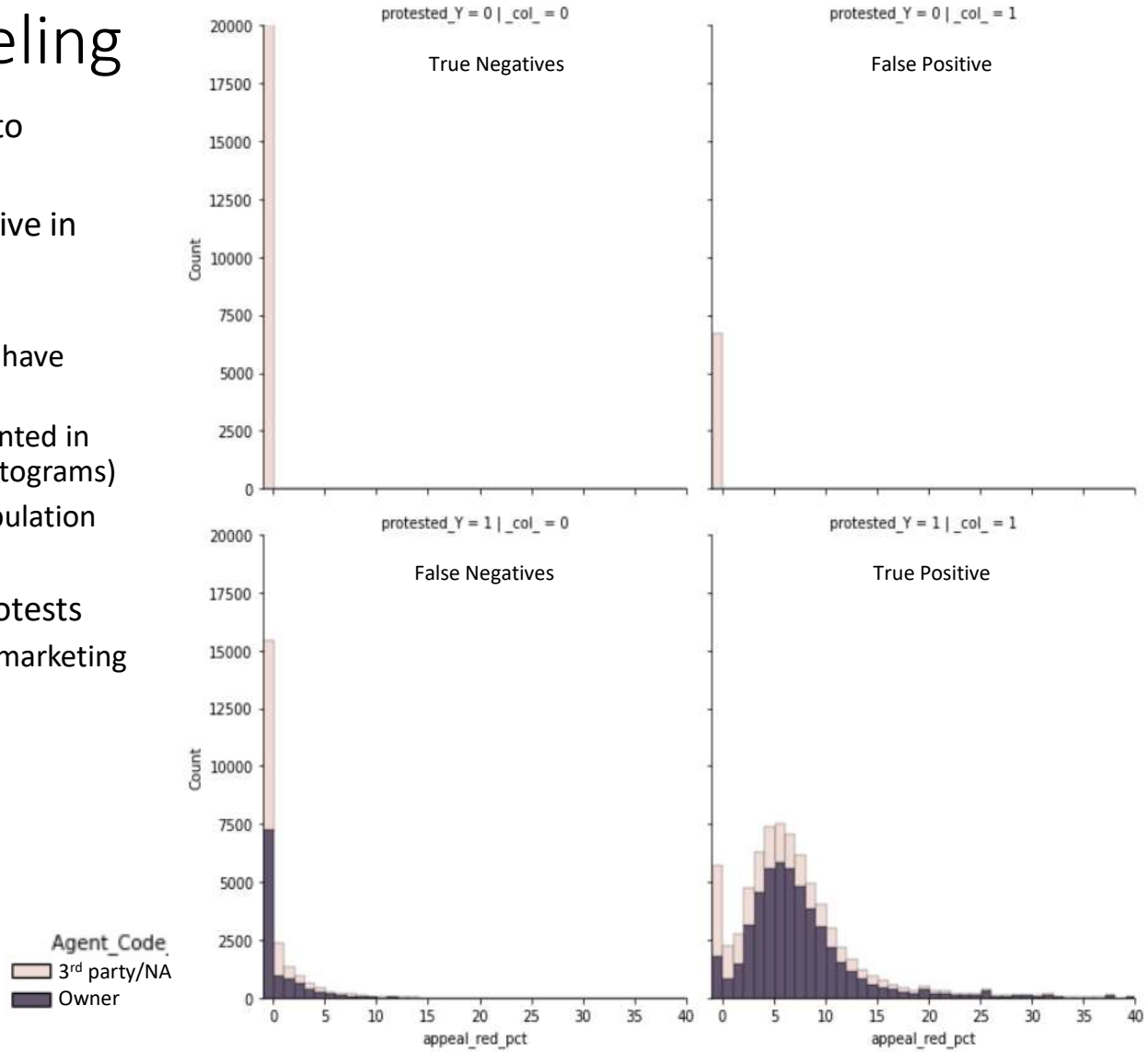
Predicted/Actual	0	1
0	246711	6719
1	22352	73622



Protest prediction modeling

Misclassified properties may provide insight into protests

- False negatives were significantly less effective in appeal reduction than true positives.
  - About ¼ of all protests fall in this category.
  - 2/3 of 0% reductions were not predicted to have been a protest
  - 3rd Party managed protests are overrepresented in False negatives (and 0% reduction bin of histograms)
  - HCAD could benefit from educating this population on process to avoid protests in the future.
- False positives would likely be successful protests
  - These properties might benefit from direct marketing from 3rd party protest managers



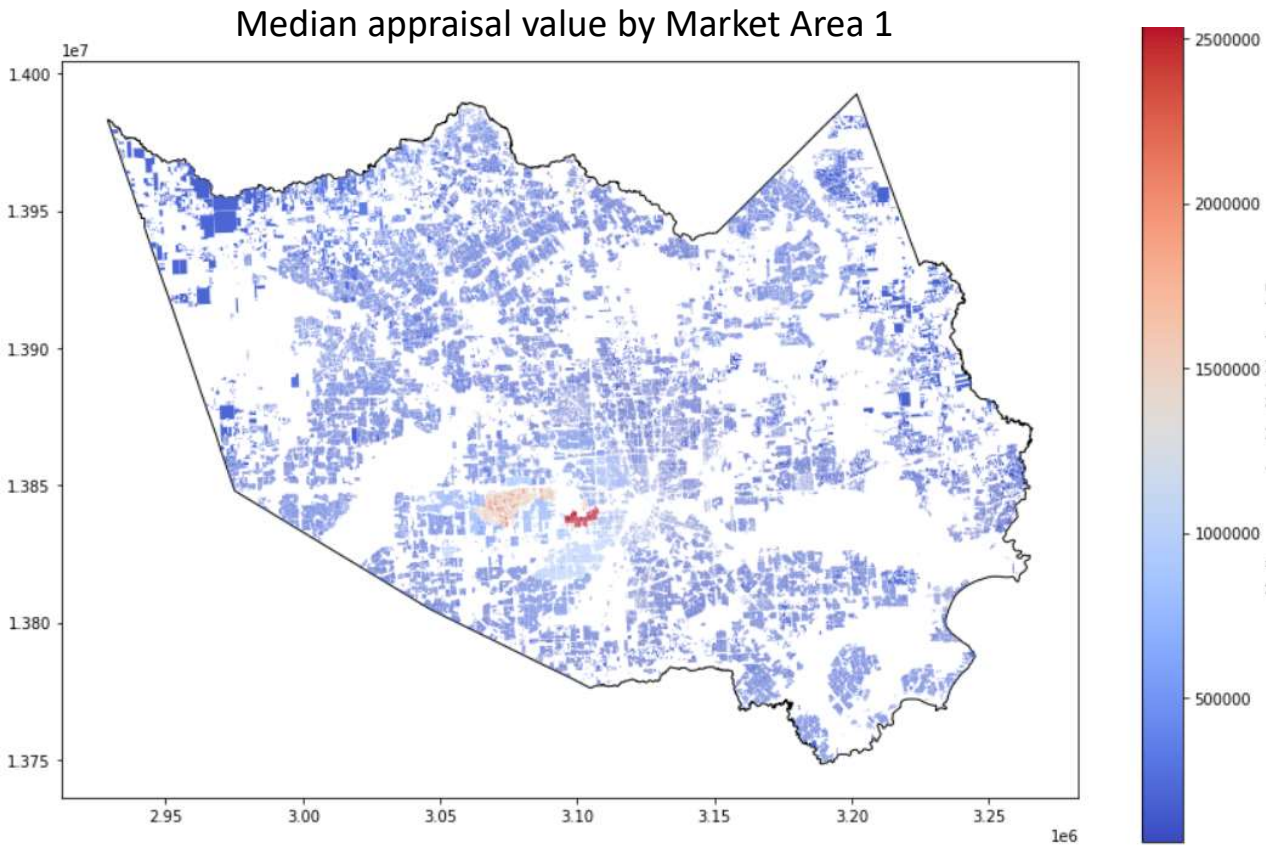
Conclusions

The 2020 HCAD data can be utilized to generate both a model for appraisal value and prediction of protest.

Both models provided insight into the appraisal process:

- Appraisal value was most impacted by building size, market area, and neighborhood
- Protests were most impacted by extra feature value, appraisal value increase, and initial appraisal value

Additional analysis into the misclassified properties could yield opportunities for 3rd party protest managers or for HCAD



Recommendations for Further Work

1) Alternative methods for handling categorical data could improve total appraisal value model.
    - The mean $/SQFT factors generated in this work were some of the most influential features in the best model
    - Analysis could be performed with alternative summarization methods
    - Analysis using an alternative feature for ratio creation (other than $/SQFT)
2) Additional data sources could be included
    - Real estate sales
    - Crime data
    - Non-residential land use data
3) These models could be applied to 2021 data to further investigate predictive power
4) Additional investigation into what prompts ineffective protests that the model failed to predict.