Capstone 2: Final Project Report
Springboard Data Science Career Path
Kevin Wolfe

Prediction of Single-Family Home Appraisal Value and Valuation Appeals

Introduction:

Every year Harris County, the county in Texas containing much of Houston's metropolitan area, reappraises the value of nearly 1.2 million residential properties (2020 Mass Appraisal Report). The appraisal values are calibrated based on sales data from the previous year in each of the counties Independent School Districts (ISDs). Despite the appraisal district's assertion that they have appropriately valued the properties, over 1⁄3 of the properties in Harris County were appealed in 2020 and the number of appeals has been growing consistently the past several years (Community Impact). Harris County residents are motivated to appeal an increase in their property values because these are used to calculate the property taxes owed. Based on appeal data 68% of these appeals are managed by consultants (Community Impact). Most of these consultants charge a fee from 20-50% of the taxes saved through appraisal reduction. Assuming the average charge of $200 per appeal (I received several flat rate offers of this price to represent me) Harris County residents likely spent $56 million fighting their appraisal value. The primary goals of this study are to: 1) validate that a model based on property description can reasonably predict appraisal values. 2) Determine if an appeal can be predicted and identify criteria that predict an appeal.

Databases:

Harris County Appraisal District (HCAD) maintains a database of all properties and their descriptive information in Harris County TX. The following databases were utilized for this analysis:

Real Property Data (https://download.hcad.org/data/CAMA/2020/Real_acct_owner.zip) Available from 2005 to 2021, 2020 data was utilized for this analysis. This file contains account and descriptive data about every property in Harris Co. this includes previous property size and previous value information.

Hearings Data (https://download.hcad.org/data/CAMA/2020/Hearing_files.zip) Available from 2005 to 2021, 2020 data was utilized for this analysis. This file contains information about each of the properties that was appealed, this includes the initial and post-hearing values.

Building Information (https://download.hcad.org/data/CAMA/2020/Real_building_land.zip) Available from 2005 to 2021, 2020 data was utilized for this analysis. This folder contains information about the building. The fixtures.txt file was specifically used to add information about each property's structure.

Tax parcels (https://download.hcad.org/data/GIS/Parcels.zip) GIS data for each parcel in Harris Co. Utilized for map-based visualizations.

Potential Clients:

1) Single Family Homeowners within Harris Co. TX – This analysis will provide homeowners a potential pathway to confirm that their property has been valued properly. In addition, post-analysis there will be properties who will have an appeal predicted who have not appealed their properties value. These homeowners may benefit by being notified of this prediction.
2) Agents who represent property owners - This work could be used to identify potential clients that are more likely to be successful. Since many agents charge based on a percentage of reduced taxes identification of these properties could lead to higher profit.
3) Harris County Appraisal District - If there are key features that seem to trigger appeals HCAD could either evaluate this metric to see if a change in how this value is calculated is merited. Alternatively, HCAD may need to target education around these features to decrease the number of appeals that require hearings.

Data Wrangling:

Real Property Data

- Initial Dataset contained 1,490,935 rows and 71 features. This dataset is a mix of homes, businesses, farms, schools, and all other property types in Harris Co.
- The index of the dataset was set to acct.
- This dataset was reduced by dropping columns like address and property owner's name that were property identifiers beyond the account number.
- New properties (properties that did not have a prior_land_val or tot_mkt_val) were excluded from this analysis since they were just built/sold which would be the properties value.
- Finally, a subset was created of Single-Family Homes (state_class = A1).

Hearing Data

- Initial Dataset contained 417,508 rows and 15 features. This dataset has an acct column that matches with the Real Property Data that was set as the index.
- Rows with missing appraisal values were dropped.
- Date related columns were dropped.
- Like with the Property Data the dataset was reduced to only Single-Family homes (state_class = A1).
- This dataset was then joined to the Real Property Dataset.
- After joining to the Real Property Dataset properties that were not protested needed data added. Since these properties did not protest the reduction in appraised value was 0. Value based columns were filled by the corresponding final appraisal values (from the Real Property data table).

Fixtures – subset of Building Information.zip

- Initial dataset contained 8,661,675 rows and 5 features. This dataset contains the acct column, but it is not indexable as presented (the same acct number may appear many times in separate columns).

- This dataset was pivoted by acct, summing the features to create a dataset of the totals for each of the possible fixture types.
- The pivoted data was then joined with the previously combined Property and Hearing datasets.

Data Exploration:

The final dataset had 1,059,015 rows and 90 feature columns. Full descriptions of all data columns can be found within the HCAD PDATA codebook (https://hcad.org/assets/uploads/pdf/pdataCodebook.pdf). Many of these columns have data only presented as codes which require the descriptions contained within (https://download.hcad.org/data/CAMA/2020/Code_description_real.zip).

tot_appr_val

The distribution of the total appraisal value can be seen below (Fig 1). This value is equivalent to the sum of land_val, bld_val, and x_feature_val (Fig 2). While the overall distribution looks close to normal there are some extremely high value properties that dramatically impact the summary statistics for this feature. There are several columns that are directly related (and correlated) to the tot_appr_val, while continuing to be utilized in this analysis because they are not identical to tot_appr_val, for brevity I will simply state that the distributions for these columns look nearly indistinguishable. These columns are: assessed_val, tot_mkt_val, prior_tot_appr_val, prior_tot_mkt_val, new_construction_val, tot_rcn_val, Initial_Appraised_Value, Initial_Market_Value, Final_Appraised_Value, Final_Market_Value.

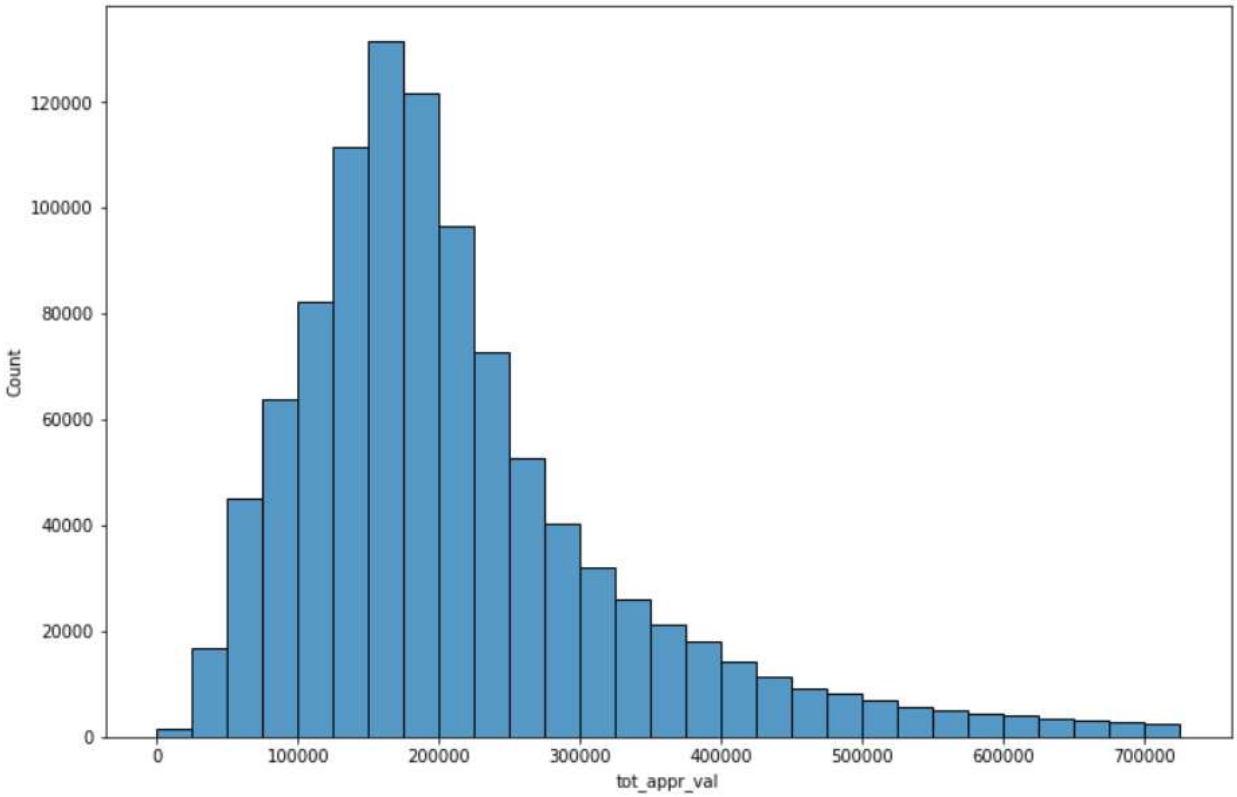
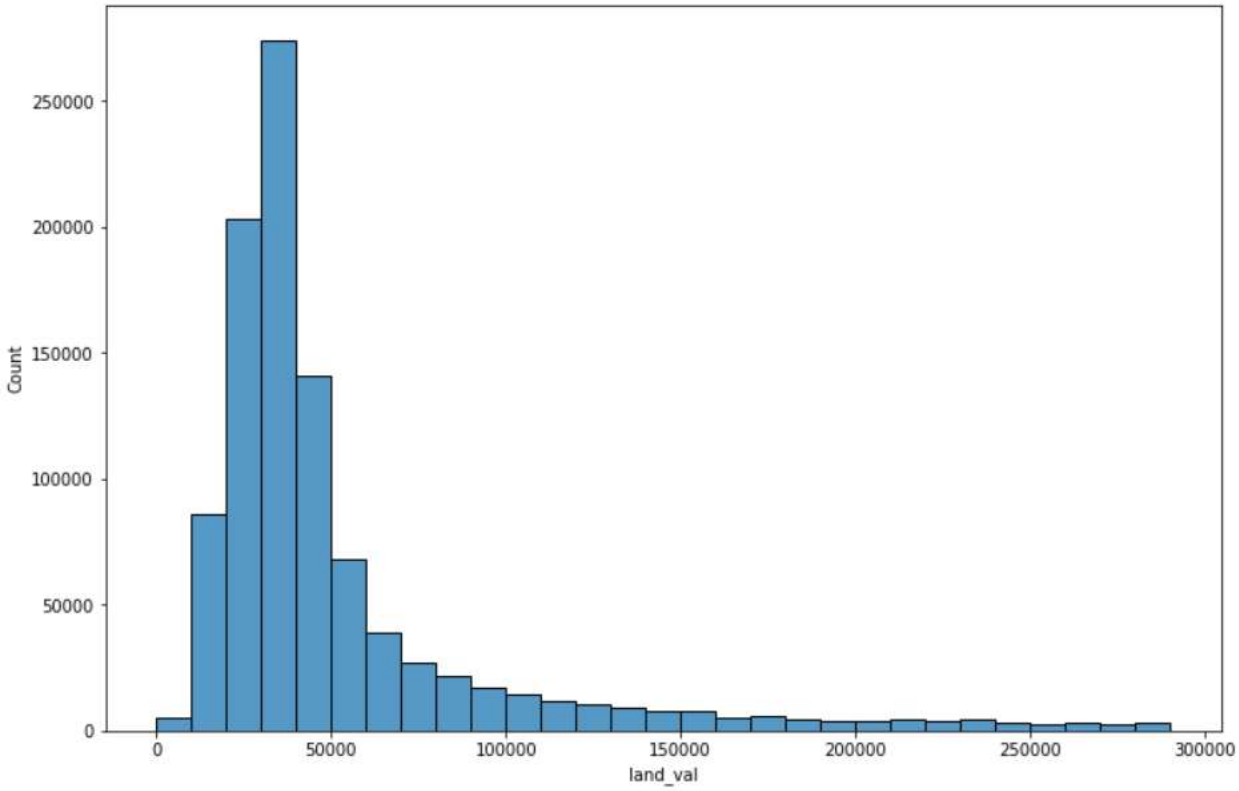
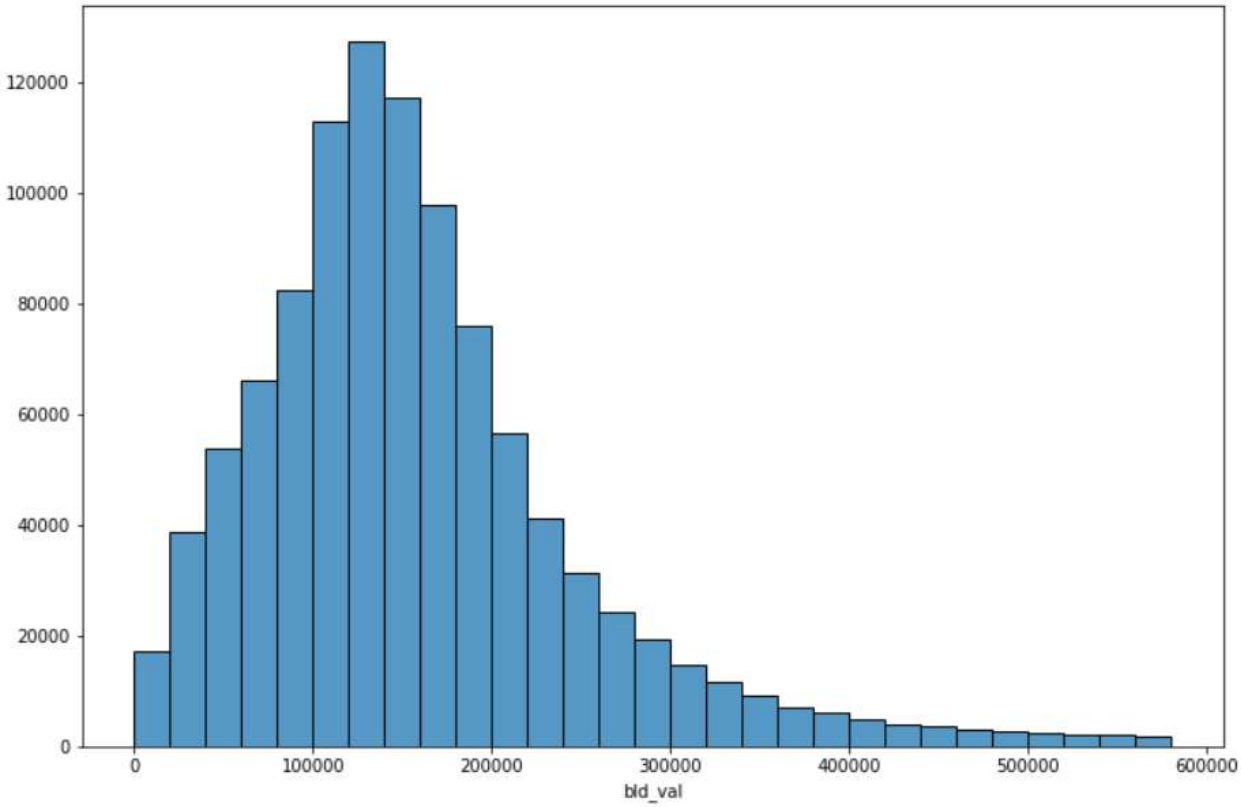


Figure 1: tot_appr_val = Total Appraisal Value for the 1.06M single-family homes within Harris Co. TX. Bin size = $25,000. Mean = $260,227, Median = $190,169, std = $318,229.


Count
250000
200000
150000
100000
50000
0
0
50000
100000
150000
200000
250000
300000
land_val


120000
100000
80000
60000
40000
20000
0
0
100000
200000
300000
400000
500000
600000
bld_val

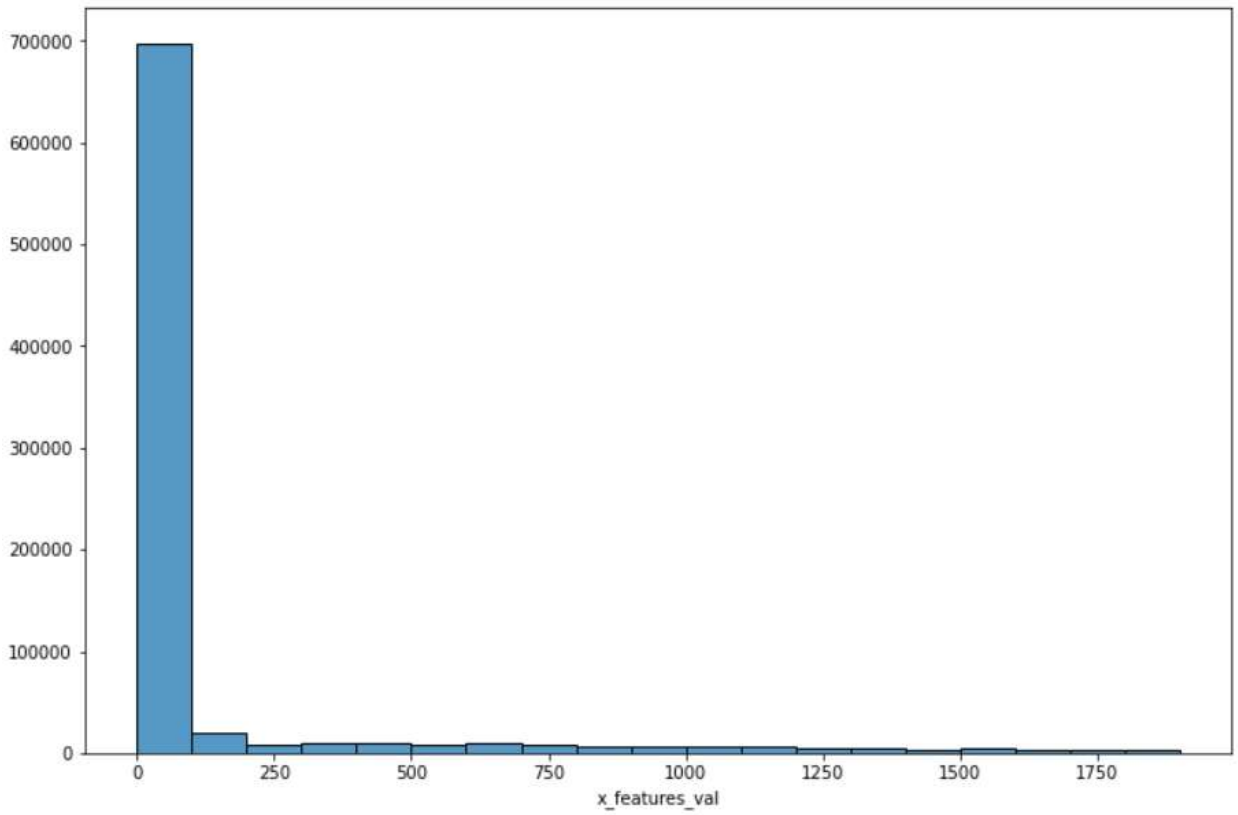


Figure 2: Distributions of land_val, bld_val, and x_features_val the three components of tot_appr_val. Note that most properties have no x_features_val. Land and building values display similar distributions to tot_appr_val. Land_val and bld_val are some of the most strongly correlated features in the dataset. Like with tot_appr_val, there are corresponding columns named prior_land_val, prior_bld_val, prior_x_features_val that look nearly identical.

Categorical Features

Categorical features presented a challenge in this dataset due to the very large number of possible categories possible for many of the features. Table 1 presents the category along with the count of unique values.

Table 1: Categorical Features that were converted to factors.

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

In addition to being a problem in machine learning due to the number of columns that would be created were one hot encoding utilized for these columns the fact that the categories were stored as integers made it necessary to either process these columns in some way or to drop them from the analysis. An attempt to transform these categorical features was performed by creating a value factor and then

mapping this value factor to the categories. This was calculated as the mean $/sqft for the unique value within the category divided by the mean $/sqft for all properties. This calculation was performed for both the land_val and bld_val columns. The calculated factors can be considered a proxy to the desirability of land and structures in a given area of Harris Co. All these factors experienced their maximum values in the area West of downtown Houston (Fig. 3).

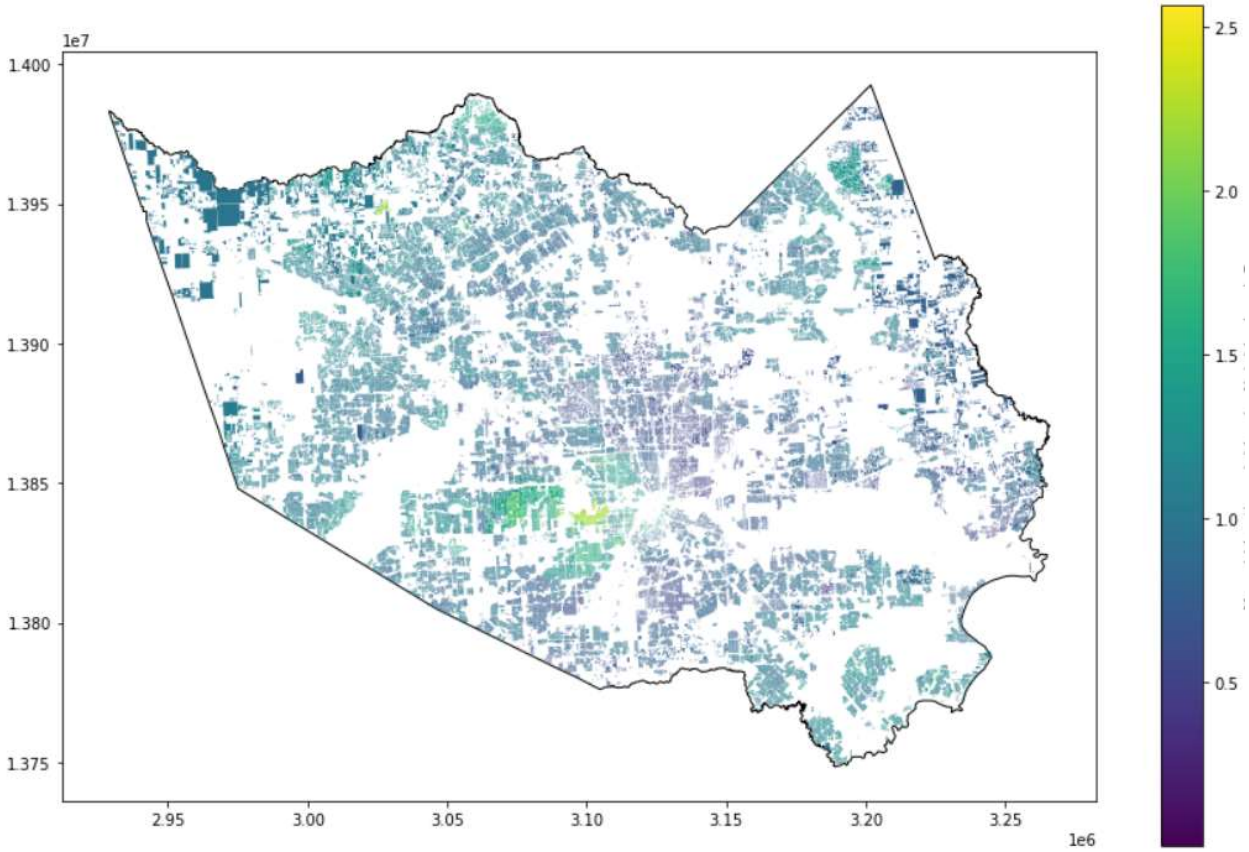


Figure 3: Example distribution of a ratio factor created to accommodate categorical columns. In this example the bld_val factor for neighborhood_Grp for within Harris Co is displayed.

protested

One of the major response variables that I intend to predict. This is the individual single-family homeowners choice based on their initial assessment. About 27% of the values are Y which means that nearly 1/3 of property assessments are review via some form of hearing (Figure 4).

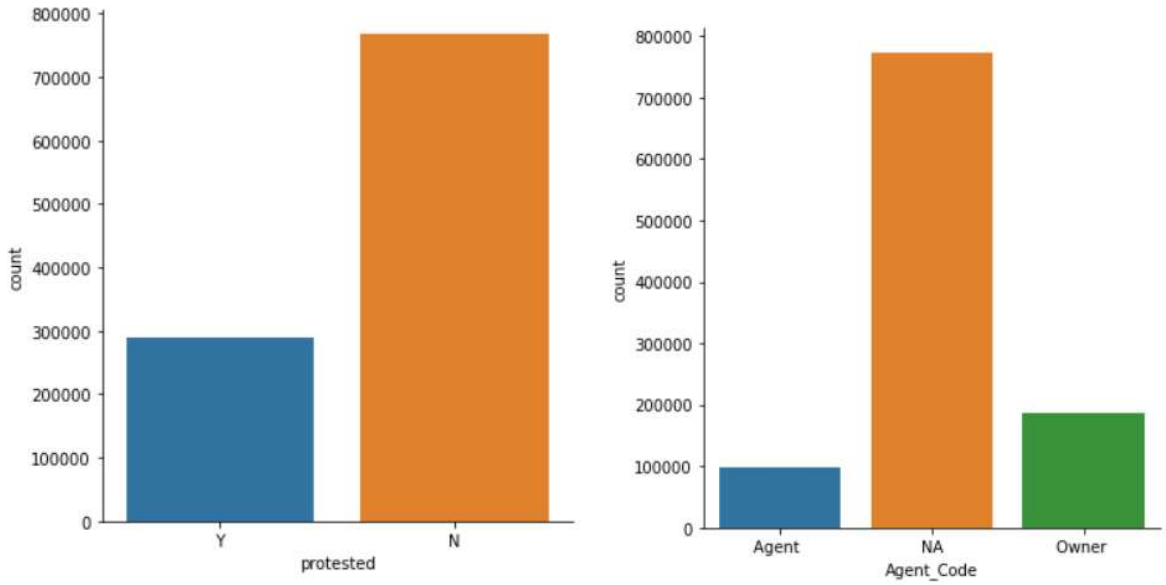


Figure 4: Distribution of categorical columns protested and Agent_Code. Protested is one of the values that this work will attempt to model. About 27% of the values for singel family homes are Y for this column indicating that a protest was filed for this property. Agent_Code is either Agent for protests managed by a 3rd party, Owner for protests managed by the Owner, or NA for properties that were not protested.

Agent_Code

For protested properties an Agent_Code has been supplied indicating who attended the hearing. Owner indicates the property owner self-represented in their protest, Agent indicates that a 3rd party represented the property owner at the protest hearing. 27% of the values are Agent or Owner because only protested properties have a representative at a hearing.

appeal_reduction

This feature was calculated as Initial_Appraised_Value minus Final_Appraised_Value. If no appeal was filed this value is $0. The distribution for only the appealed properties is presented below. It is interesting to note that 1/3 of the appeals represented by an agent saw <$2500 reduction in appraised value.

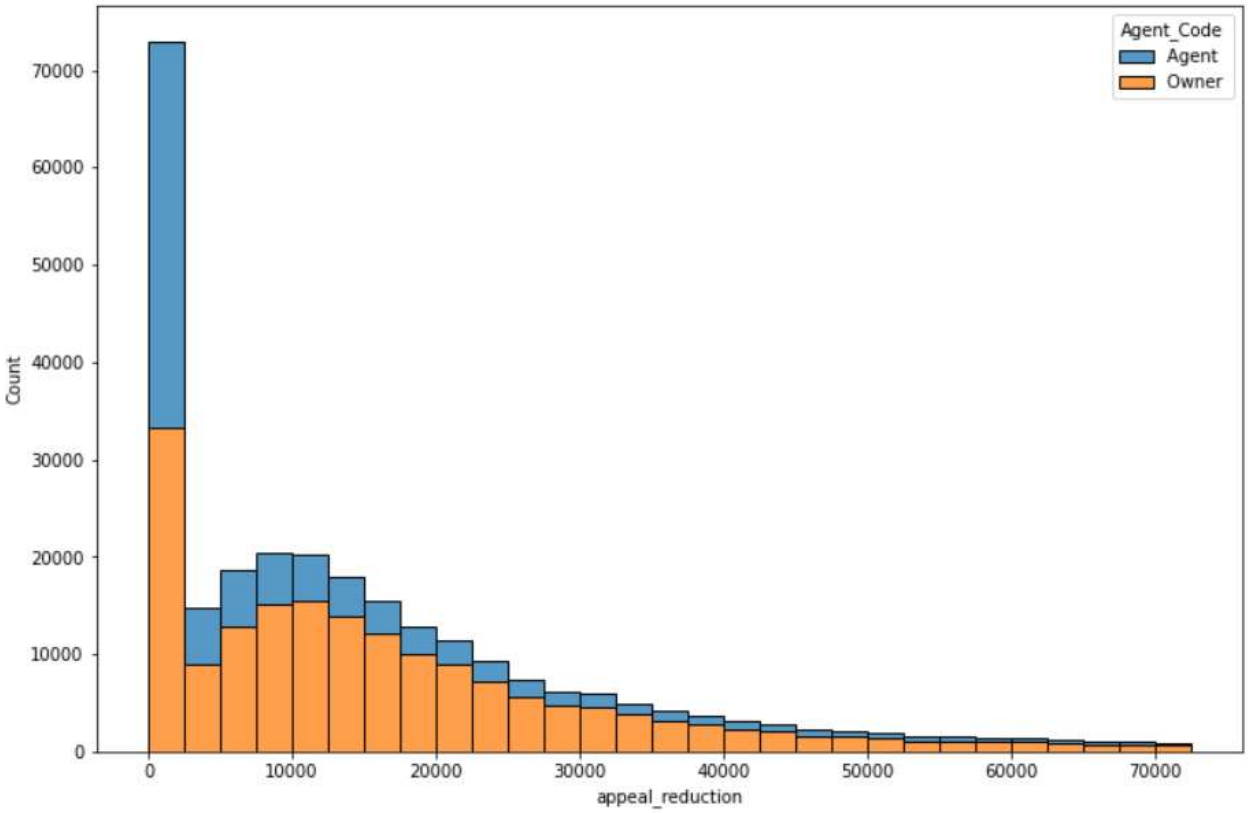


Figure 5: appeal_reduction distribution colored by Agent_Code. 25% of appeals resulted in less than a $2500 reduction in value. Over half of these ineffective appeals were managed by a 3rd party.

Data preparation

In preparation for modeling datasets were split into 67% training and 33% testing subsets. Prior to this split, remaining categorical columns were converted to one-hot encoding for compatibility with models. After the split, a StandardScaler was trained on the training dataset and applied to numeric columns in both the training and testing datasets.

Data Modeling

Total Appraisal Value Prediction

HCAD data from 2020 was split into two groups. 67% of the original dataset was used in creation of a series of models to predict tot_appr_val. The remaining 33% was held aside for model testing and evaluation. The dataset utilized in this analysis attempted to minimize the usage of value-based features and excluded all information from the protest table since these columns were almost exclusively value related. Five models were created in the attempt to predict appraisal value based on HCAD data. The initial 3 models (one linear, two ridge models) performed at very similar levels based on both Pearson's R-squared score and root mean squared error. A final linear regression that limited modeling to the k best columns was created which was a dramatic improvement in RMSE despite a lower Pearson's score

(Table 2). All 4 of these models perform poorly against the training data for the most expensive properties in Harris Co. (Figure 6).

Table 2: Appraisal Value Models, score, and RMSE

Model Description	Training Score	Test RMSE
Linear Regression - default	0.78	4352300
Ridge Regression - default	0.78	4352357
Ridge Regression – alpha = 10	0.78	4352868
Linear Regression – k best columns	0.72	167434
Random Forest Regression – limited column number	0.99	50404

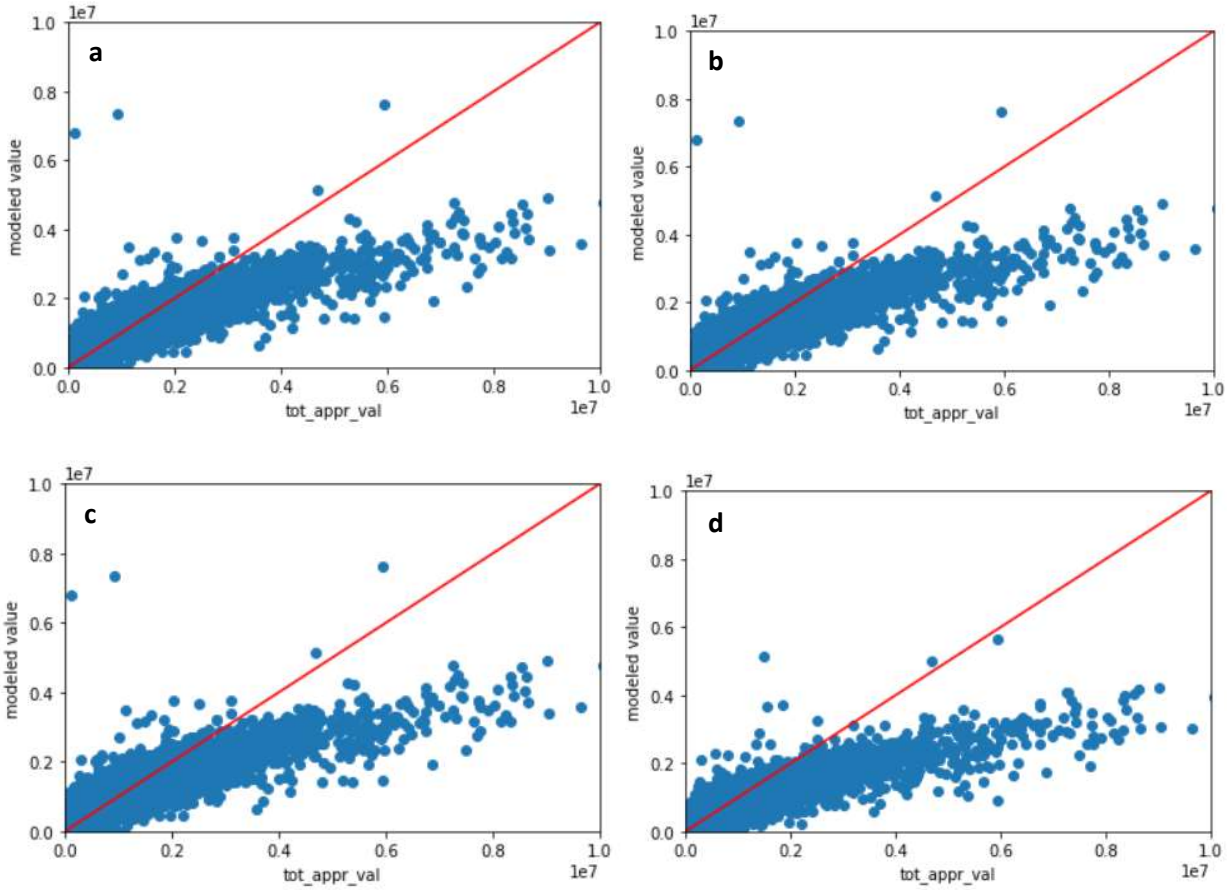


Figure 6: training performance modeled tot_appr_val vs. actual tot_appr_val for Initial 4 models. a = Linear Regression – default, b = Ridge Regression – default, c = Ridge Regression – alpha = 10, d = Linear Regression – k best columns. Red line indicates 1:1 ratio.

The final model created for this analysis utilized a Random Forest Regressor. Due to computational limitations associated with the number of rows and columns in this dataset the 9 most influential features determined by the selectkbest__k attribute available through GridSearchCV. While there was no rollover where additional features did not improve the score it is visually apparent that the first 9

features provide the most uplift (Figure 7). The retained features were: bld_ar, FXT, land_factor_Neighborhood_Grp, bld_val_per_ft2, bld_factor_Market_Area_1, land_factor_Market_Area_1, land_factor_school_dist, land_factor_Market_Area_2, land_ar.

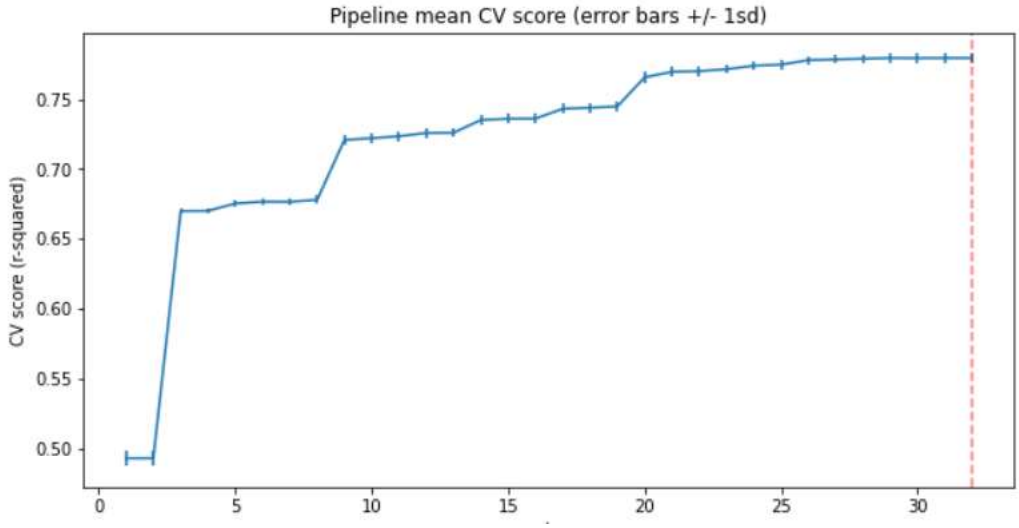


Figure 7: CV score as each additional feature is added. Based on this result it was decided to keep at least the top 9 features.

The resultant Random Forest Regressor was by far the best model produced for tot_appr_val. Unlike with the previous models, more expensive properties were successfully predicted which yielded significantly better scores. Comparison of the training and testing fit is presented in figure 8.

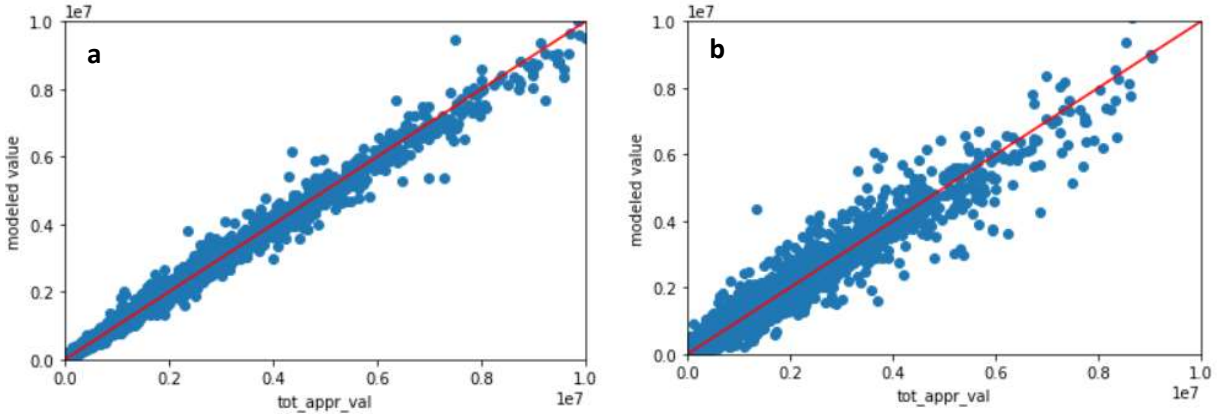


Figure 8: tot_appr_val vs. model predicted tot_appr_val for a) Training Data and b) Testing Data for the Random Forest Regression model. 1:1 line in red.

Protest Prediction

The second group of models created in this analysis were to evaluate if a protest can be predicted for single-family homes in Harris Co. Like with the total appraisal value prediction 67% of the data was used for model building and 33% for testing. Unlike with the total appraisal value prediction models, value-based information was included in the dataset for these models. Modeling was performed using decision trees and random forest models, yielding 4 potential models (Table 3). Performance was judged based on the Precision and Recall for "Yes" prediction as the properties who do protest (or at least are predicted to protest) are of interest to the potential clients for this study.

Decision Trees and Random Forest Classifiers were selected in this analysis because these methods are well suited for predicting a yes/no variable like did a protest occur. Both default classifiers performed better due than the GridSearch versions because the hyperparameters had a tendency to decrease the overall depth of the resultant tree. The GridSearch Decision Tree was negatively impacted by this more than the Random Forest GridSearch. Both of these models were restricted in the number of hyperparameters that could be tested due to the extremely long run times (often >3 hrs) for these models.

Table 3: Comparison of models created for prediction of protests. Random forest models performed the best for this prediction.

Model	Accuracy	Precision - Yes	Recall - Yes
Entropy Decision Tree – default hyperparameters	0.85	0.74	0.74
CV Decision Tree – criterion = gini, max depth = 8, min_samples_leaf = 0.01	0.82	0.74	0.52
Random Forest Classifier – default hyperparameters	0.92	0.92	0.77
CV Random Forest – max_samples = 100,000, min_samples_split = 2, n_estimators = 100	0.90	0.90	0.72

The best models (Random Forest Classifier) performed better at not predicting protests where one did not occur than correctly predicting the actual protests (Precision > Recall). On further investigation of the protested properties that were misclassified (false negatives) it appears that these properties were less successful based on appraisal reduction % in their protest than the true positives (Figure 9). An additional observation with this model is that a large proportion of the false negatives were protests managed by a 3rd party. This may be a result of many of the offers from 3rd party agents are "no risk" offers which may prompt protests that are not merited.

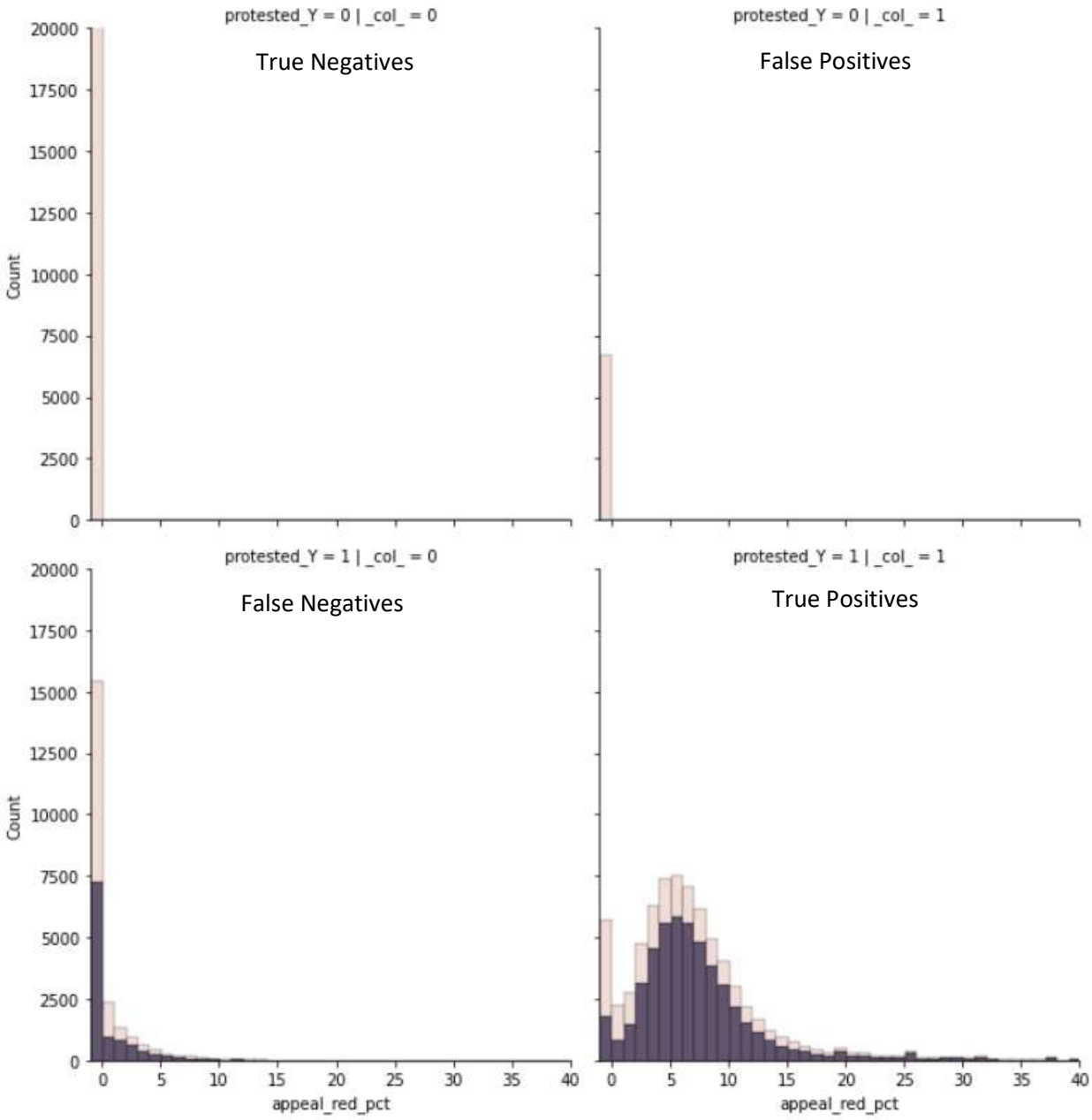


Figure 9: Predicted vs. Actual protest properties. This has been charted against the reduction in appraisal amount (%) and colored by the representation present at the appeal: owner = grey, 3rd party or NA = tan.

Potential Utilization:

1) Utilization of the predicted total appraisal value in a protest. This would need to be evaluated for the individual property (and the HCAD provided offsets) to be of use.
2) The protest prediction model could be used to identify properties which are likely to be successful in a protest. The false positives from the 2020 dataset would be a group that could be

worthwhile for a 3rd party protest representative to directly market their services to. In addition, the 2020 model could be used on the 2021 HCAD initial dataset to identify a list of properties expected to protest.

3) The protest prediction model could also be used by HCAD to identify properties which are protesting but are less successful. Potentially the false negatives could be provided with education materials to decrease the total number of protests HCAD has to process.

Further Work:

- Like with almost all predictive models, additional tuning could be attempted for the hyperparameters for the final models presented with this work. In addition to additional hyperparameter tuning, alternative methods for creating the factors for the categorical data could be investigated. The method utilized in this work used the mean $/sqft, however median $/sqft may be worth investigation.
- Additional data sources like real estate, crime, or other land use information in the immediate area could be used to provide additional data for better constraining the tot_appr_val prediction.
- 2021 initial appraisals have been released at this time. A further testing of the models created with 2020 data presented here would be to predict the properties that will protest in 2021. This could be checked against the 2021 hearing results once they are released to evaluate effectiveness.

References:

2020 Mass Appraisal Report, https://hcad.org/assets/uploads/pdf/Reports/2020-Mass-Appraisal-Report.pdf, Accessed 6/1/21

Community Impact, https://communityimpact.com/houston/katy/housing-real-state/2021/05/20/more-than-ever-harris-countys-appraisals-draw-protests/, Accessed 6/1/21

Property Downloads (hcad.org), https://hcad.org/pdata/pdata-property-downloads.html#, Accessed 6/1/21