Project Report:
Predict Future Home Value Index

Overview

The team set out to develop a predictive model for the short-term change in home price value in metro areas across the United States.

Market data from Zillow, job opening statistics from the Bureau of Labor Statistics (BLS), and Gross Domestic Product (GDP), population and personal income statistics from the Bureau of Economic Analysis (BEA) were utilized to train and test a variety of models.

The final Random Forest Regression model was able to use 3.5 years of training data across 471 metro areas to predict the home value percentage change with a mean absolute error of .24% (less than 40% of one standard deviation) on the validation dataset covering July 2022 – December 2023. Additionally, the model correctly predicted a decline in home value 90.4% of the time, when there was a true decline in home value. The model correctly predicted an increase in home value 83.3% of the time, when there was a true increase in home value.

Problem

Many home buyers and sellers are considering entering the market and want to understand how they should expect prices to change from the time they begin researching the market until they are ready to transact.

Having a short-term feel for where the market is going would help both prospective home buyers and sellers to determine how urgently they want to enter the market and what changes they can expect over the next month.

Home values must be predicted for the next month. Zillow publishes the Zillow Home Value Index (HVI) measure each month. This measures the typical home value across a given region reflecting the typical value for homes in the 35th to 65th percentile range. This is the target measure used for our analysis to predict short-term changes in home values

Approach

Data wrangling

Zillow

Zillow provides a variety of datasets with relevant market data beyond just the HVI. This data includes market relevant statistics such as rent, for sale inventory, percentage of housing sold above listing, median days to close and new constructions. The metrics are available at a variety of levels, ranging from zip code to metro (i.e. city and surrounding area) to state level. The team determined to perform the analysis at the metro level, since that was the lowest level with a significant amount of statistics available. Zip code level statistics were ignored. Each metro area was tagged to one state, and therefore state level statistics were easily merged onto the metro statistics by state.

While data was available going back to 2000, most key statistics were only available beginning in 2018. All data was kept for further analysis, and this observation was used to determine appropriate testing and validation datasets.

BLS

The Bureau of Labor Statistics provides monthly statistics for the number of job openings in each state, going back to 2004. Seasonally adjusted and raw statistics were both imported for further analysis.

BEA

The Bureau of Economic Analysis provides quarterly statistics for personal income, per capita personal income, population, and current dollar GDP by state going back to 2000. All statistics were imported for further analysis.

Exploratory data analysis and preprocessing

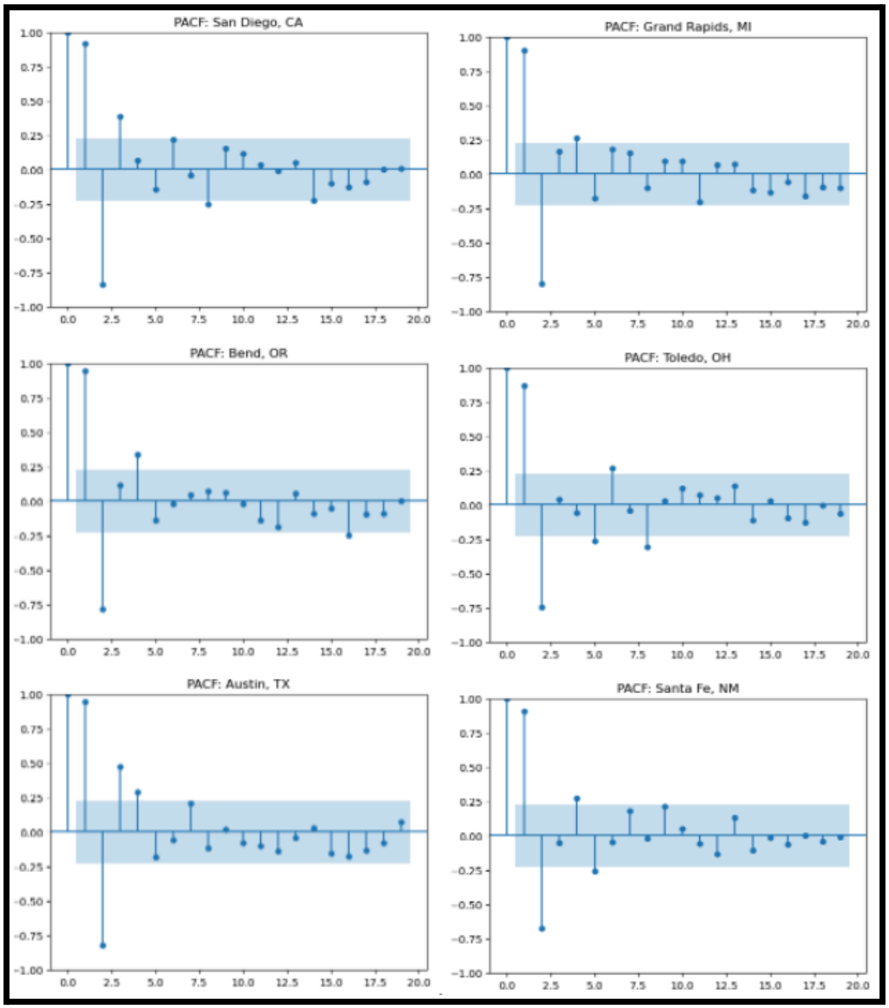
The first step in the exploratory data analysis was to aggregate all source data into one dataset. The datasets where joined on state to create one observation for each metro per period with all statistics. The resulting dataset included 16 independent variables:

metro_hvi
metro_for_sale_inventory
metro_med_days_to_close
metro_new_construct
metro_new_listing
metro_pct_abv_list
metro_pct_blw_list
metro_pct_w_pricecut
metro_rent

state_hvi
state_personal_income
state_personal_income_per_capita
state_population
state_gdp_cur_dol
state_job_openings
state_job_openings_szn_adjd

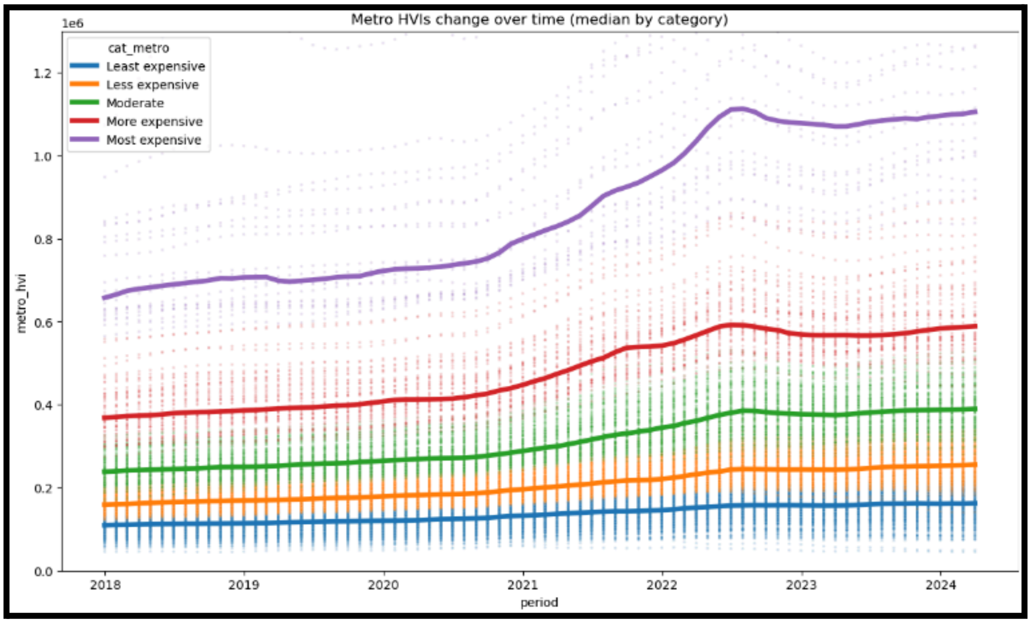
As the objective of this model is to predict future home prices, additional features were needed to assess the change in values over time. Rather than looking at the metro_hvi as the dependent variable, the change in metro_hvi between the current and subsequent period was added to the dataset as the dependent variable (“frwd01_mon_metro_hvi_pct_chg”).

Additional lag features were created to supplement the existing features. To determine the most relevant periods to include for lag periods, a time correlation analysis was performed. The percentage change in HVI over time was analyzed using the partial autocorrelation function to measure the correlation between observations separated by k periods. A spot check across various metro areas in each expense bucket informed the decision for the creation of additional lag features.

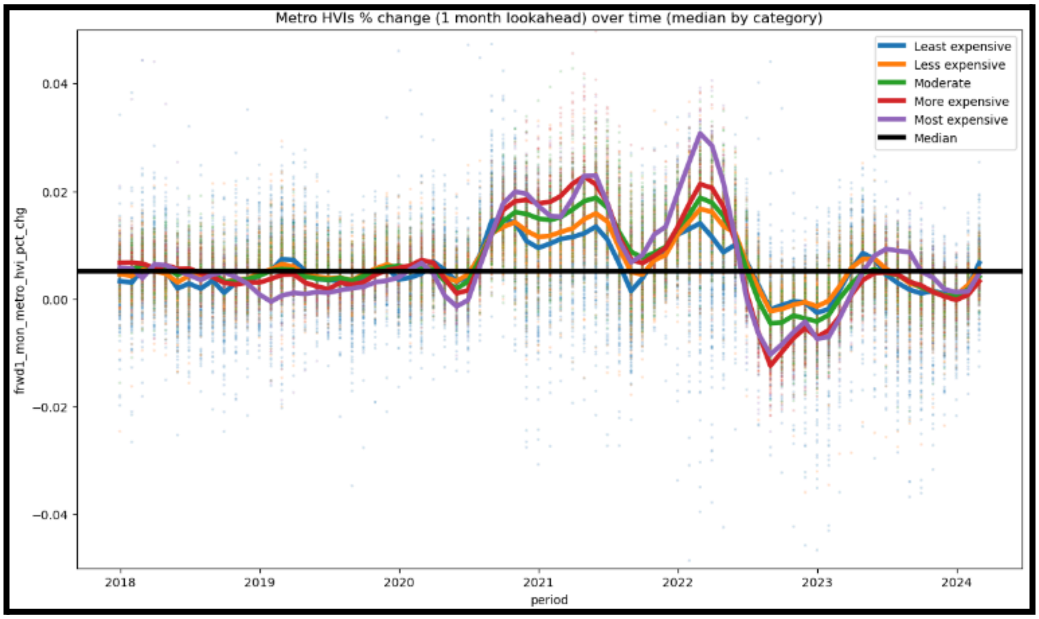


The independent variables from the previous 1, 2, 3 and 6 months were used to create additional features as both raw amounts and percentage changes. This resulted in an additional 128 independent variables.

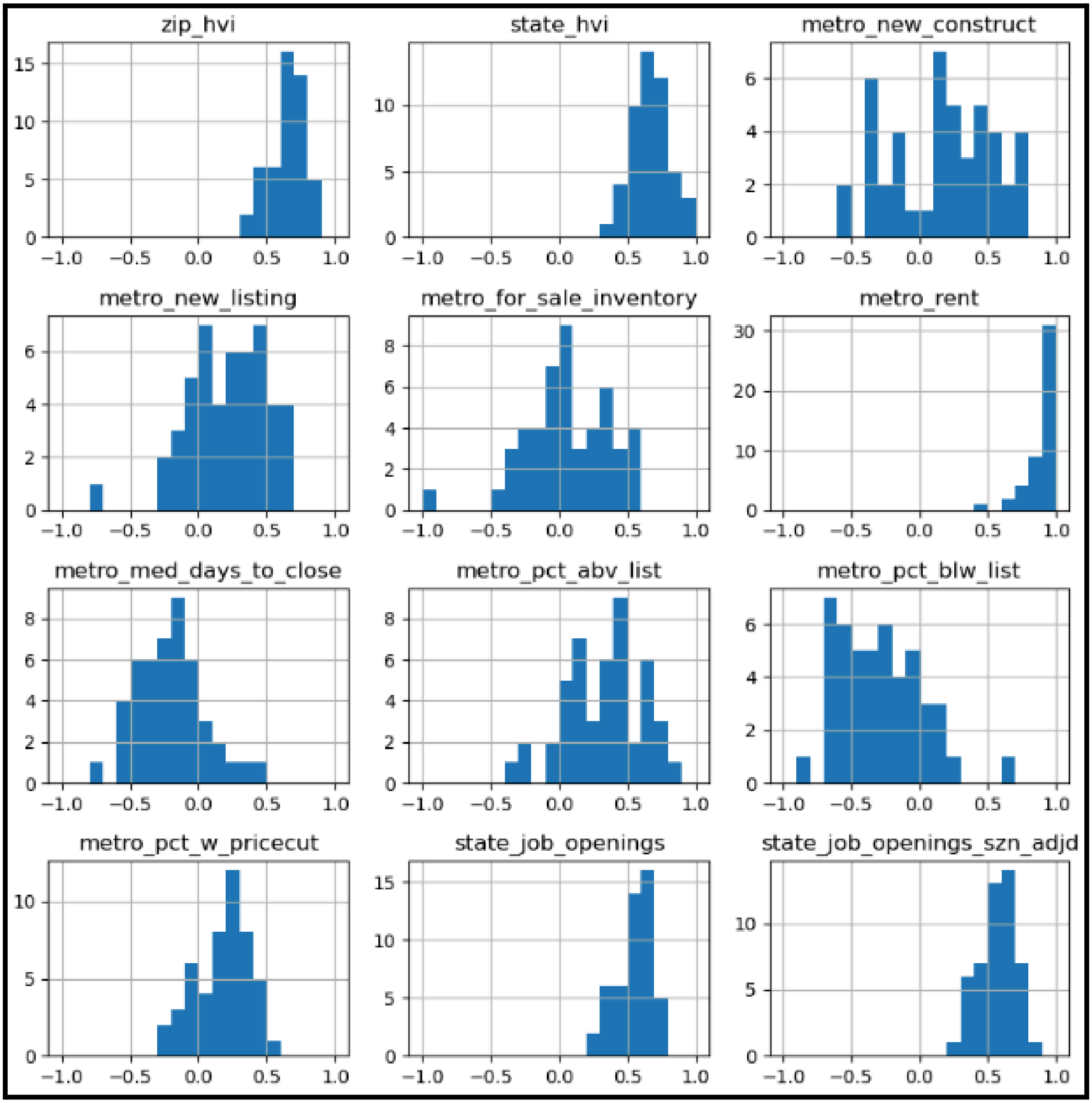
With all lag features created, the metro's were bucketed into 5 categories based on the mean HVI's. In review, it was clear there is a general upward trend in prices regardless of expensiveness category. The changes of the least expensive categories seem minimal in comparison with the most expensive categories.



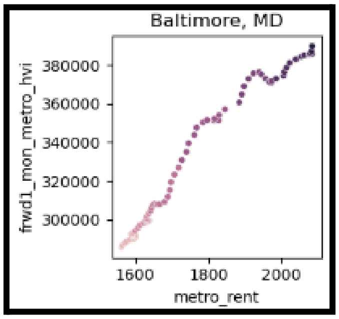
When reviewing just the percentage change, it is apparent all metro area's generally move together directionally, with varying levels of magnitude depending on the period. Is most instances the more expensive groupings will have more severe changes in price.

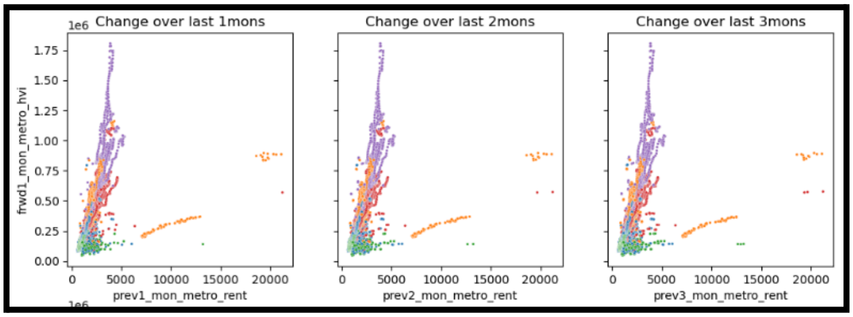


Looking at each state separately, the correlation between each independent variable and metro HVI was measured using the correlation coefficient. Looking at a histogram of the results for each state; metro rent and state job openings appear to be strong candidates for predictive value. New constructions also appear to present potential to be a valuable statistic.

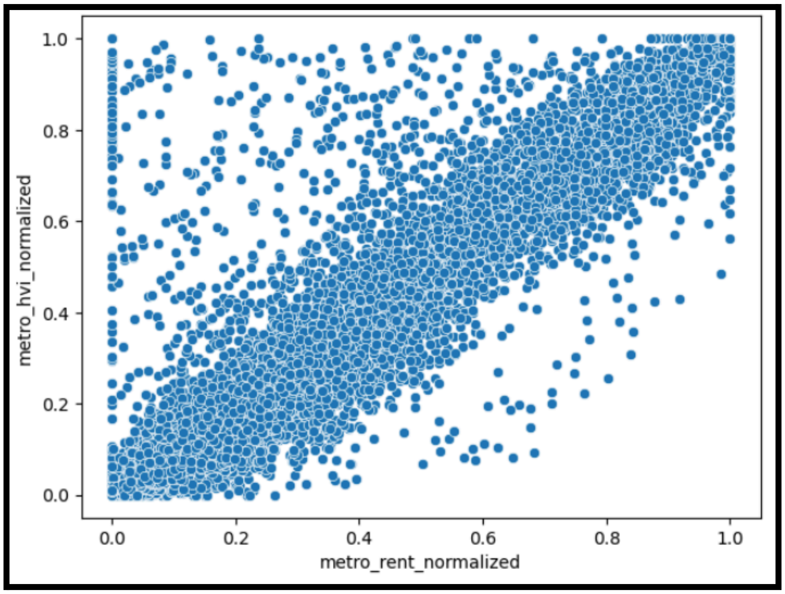


No strong linear relationships were identified in aggregate for the population. Looking at the dataset as a whole, there were loose linear relationships identified. To find stronger relationships, the data needed to be viewed separately for each metro or state. For example, it is clear when looking at Baltimore, MD rent compared to home value, as rent increases, home value increases. However, when looking at all states' rent compared to home values, there is no observable relationship (primarily due to scaling).



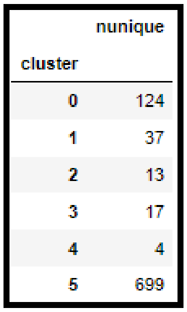


Normalization of the data provides a solution for this issue. When looking at the rent compared to the home value for all metro areas, a strong linear relationship appears in the normalized dataset. To normalize, a minmax scaler was used on the initial independent variables to scale all amounts for each metro area from 0 to 1, retaining the variance of their initial shape. Linear relationships are seen in the normalized personal income, job openings, and sales below listing. Rent provides the strongest linear relationship of all variables.



Before training any models, additional preprocessing steps were taken. Any periods with missing independent variables were forward filled using the most recent value from the previous period. Any periods without earlier data available were left empty.
Finally, an unsupervised kmeans model was used to cluster metro id's for potential dimension reduction. The kmeans model used the following independent variables to classify the metro areas:

metro_hvi,
prev06_mon_metro_hvi_pct_chg,
metro_new_construct,
prev06_mon_metro_rent_pct_chg,
state_hvi,
state_job_openings,
state_personal_income,
state_population

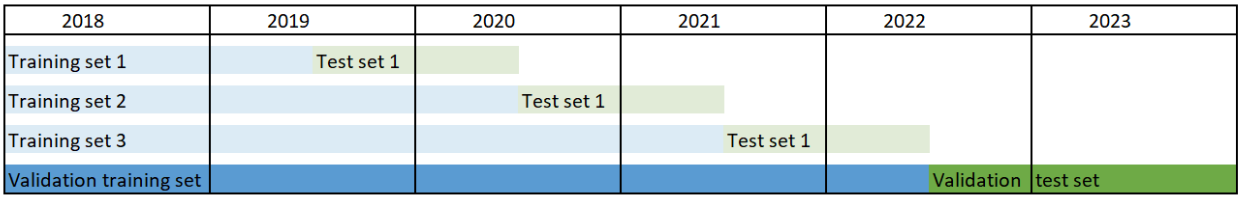


	nunique
cluster	
0	124
1	37
2	13
3	17
4	4
5	699

These variables were selected based on the results of earlier exploratory data analysis due to their stronger correlation to price. In instances where the metro area was classified into multiple buckets for different periods, the bucket with the most instances for that metro area was selected.

Model selection

The processed data set was broken into 3 training sets to be used for cross validation testing as shown in the graphic below. The final 30% of data was saved for the validation test set.



2018	2019	2020	2021	2022	2023
Training set 1	Test set 1				
Training set 2		Test set 1			
Training set 3			Test set 1		
Validation training set				Validation	test set

Mean absolute error was used as the primary metric for evaluation due to its simplistic nature. It is more resilient to outliers and provides an amount in the same units as the response variable. Additionally consideration was given to the direction of the prediction. Predictions were classified as predicting either a decrease (positive) or increase (negative). Additional consideration was given to recall and precision to evaluate how often the model would predict a decrease when the price actually declined and vice versa for actual increases.

Further, 3 separate sets of predictor variables were created for additional hypertuning.

Set 0 includes all features engineered throughout the process.

Set 1 includes all normalized amounts from the current period, 1 month lag features and all lag periods for hvi, rent, job opening change percent, population change percent and personal income change percent.

Set 2 includes all normalized amounts from the current period and all percentage change lag amounts.

Baseline

The first model was a very simple estimator, using the mean change in home value for each metro area as the predicted value.

Historical Gradient Boosting

The next model used was a Gradient Boosting Model (GBM). The HistGradientBoostingRegressor from sklearn was selected because of the size of the dataset. The model implements gradient boosting on histograms. It is an optimized implementation of gradient boosting that uses histograms to speed up the training process, particularly useful for large datasets.

This model has with native support for missing values. Therefore, no cleaning was needed and the full dataset could be run through the model. However, because records with null values were removed for other models tested, those records were also removed in the testing of this model as well.

Several variations of the model were assessed using cross-validation and random hypertuning of a few key variables. 15 iterations of variables were tested using the 3 cross validation sets for each iteration.

Learning rate was set between a range of .01 and .15.
Minimum samples per leaf were set between a range of 100 and 200.
Max depth was set between a range of 5 and 20.

The mean metrics from each of the 3 sets were summarized to assess the results of each iteration. All iterations were tested for each set of features.

Random Forest

The final model assessed was a Random Forest Model (RF). The RandomForestRegressor from sklearn was selected due to its proven success with time series analysis predictions. It is an ensemble regression algorithm. It builds multiple decision trees during training and averages their predictions to improve accuracy and robustness.

The model cannot handle null values. Any features missing more than 70% of the values were dropped completely. Any observations in the remaining dataset with a null value were dropped.

Several variations of the model were assessed using cross-validation and random hypertuning of a few key variables. 15 iterations of variables were tested using the 3 cross validation sets for each iteration.

Number of estimators was set between a range of 10 and 100.
Minimum samples per leaf were set between a range of 20 and 100.
Max features were set to either no maximum, the square root of all features or log squared of all features,.

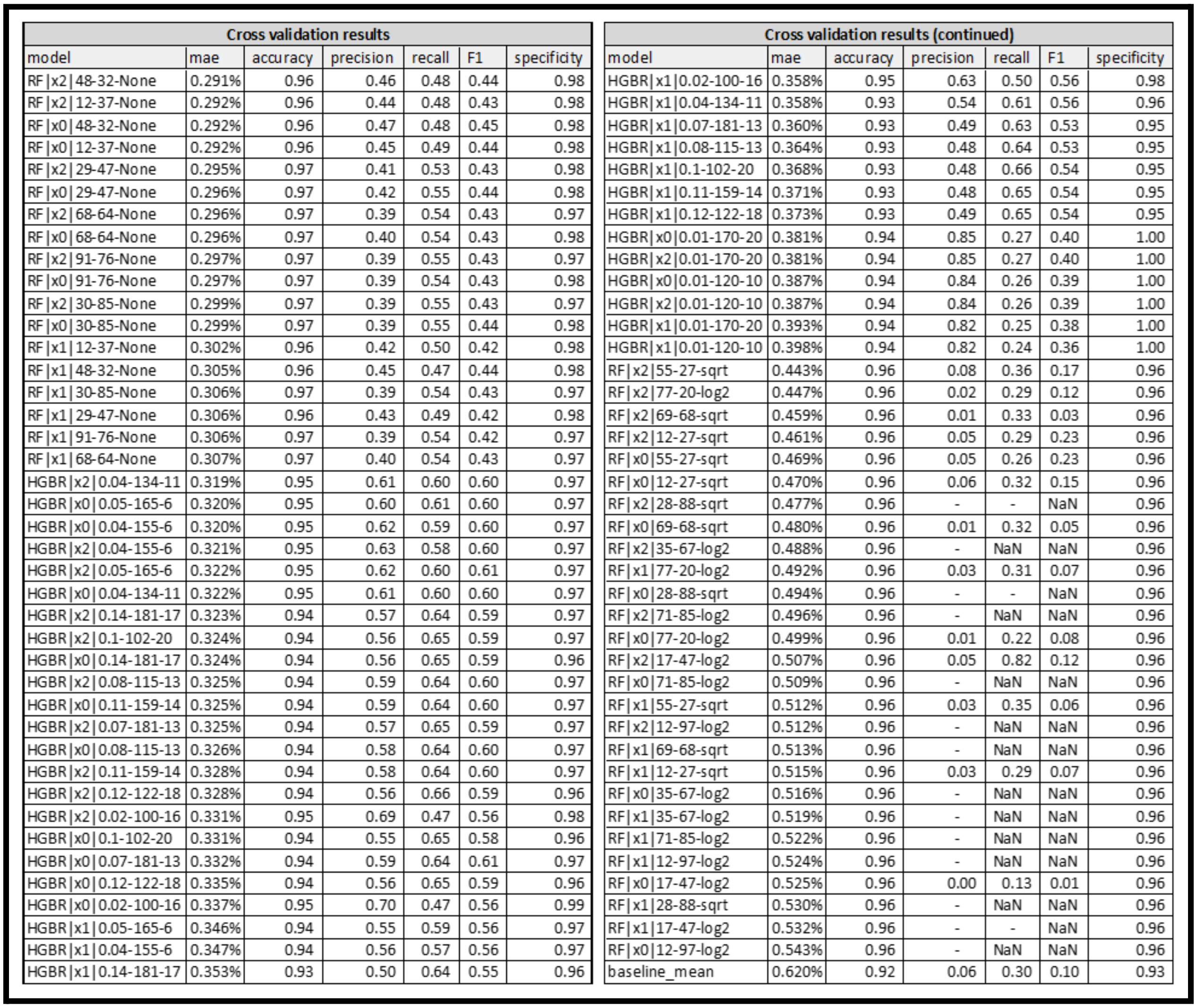
The mean metrics from each of the 3 sets were summarized to assess the results of each iteration. All iterations were tested for each set of features.

Findings

Model results

All iterations of both the Random Forest and Gradient Boosting exceeded the simple baseline. The winning model based on MAE was a Random Forest using the following parameters:

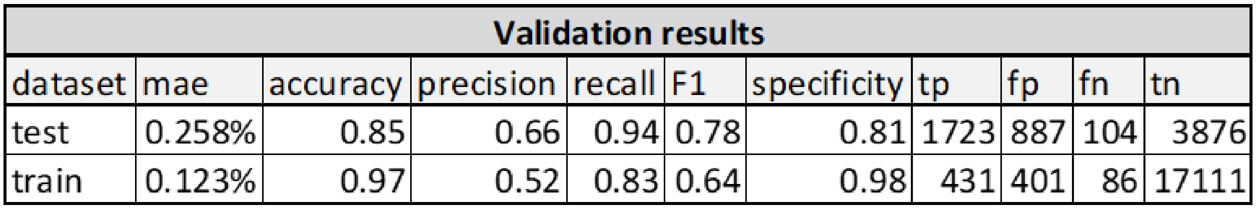
Variable set = 2 (all pct change lag and normalized amounts for current period)
Number of estimators = 48
Minimum samples per leaf = 32
Maximum features = None



Cross validation results						
model	mae	accuracy	precision	recall	F1	specificity
RF |x2|48-32-None	0.291%	0.96	0.46	0.48	0.44	0.98
RF |x2|12-37-None	0.292%	0.96	0.44	0.48	0.43	0.98
RF |x0|48-32-None	0.292%	0.96	0.47	0.48	0.45	0.98
RF |x0|12-37-None	0.292%	0.96	0.45	0.49	0.44	0.98
RF |x2|29-47-None	0.295%	0.97	0.41	0.53	0.43	0.98
RF |x0|29-47-None	0.296%	0.97	0.42	0.55	0.44	0.98
RF |x2|68-64-None	0.296%	0.97	0.39	0.54	0.43	0.97
RF |x0|68-64-None	0.296%	0.97	0.40	0.54	0.43	0.98
RF |x2|91-76-None	0.297%	0.97	0.39	0.55	0.43	0.97
RF |x0|91-76-None	0.297%	0.97	0.39	0.54	0.43	0.98
RF |x2|30-85-None	0.299%	0.97	0.39	0.55	0.43	0.97
RF |x0|30-85-None	0.299%	0.97	0.39	0.55	0.44	0.98
RF |x1|12-37-None	0.302%	0.96	0.42	0.50	0.42	0.98
RF |x1|48-32-None	0.305%	0.96	0.45	0.47	0.44	0.98
RF |x1|30-85-None	0.306%	0.97	0.39	0.54	0.43	0.97
RF |x1|29-47-None	0.306%	0.96	0.43	0.49	0.42	0.98
RF |x1|91-76-None	0.306%	0.97	0.39	0.54	0.42	0.97
RF |x1|68-64-None	0.307%	0.97	0.40	0.54	0.43	0.97
HGBR|x2|0.04-134-11	0.319%	0.95	0.61	0.60	0.60	0.97
HGBR|x0|0.05-165-6	0.320%	0.95	0.60	0.61	0.60	0.97
HGBR|x0|0.04-155-6	0.320%	0.95	0.62	0.59	0.60	0.97
HGBR|x2|0.04-155-6	0.321%	0.95	0.63	0.58	0.60	0.97
HGBR|x2|0.05-165-6	0.322%	0.95	0.62	0.60	0.61	0.97
HGBR|x0|0.04-134-11	0.322%	0.95	0.61	0.60	0.60	0.97
HGBR|x2|0.14-181-17	0.323%	0.94	0.57	0.64	0.59	0.97
HGBR|x2|0.1-102-20	0.324%	0.94	0.56	0.65	0.59	0.97
HGBR|x0|0.14-181-17	0.324%	0.94	0.56	0.65	0.59	0.96
HGBR|x2|0.08-115-13	0.325%	0.94	0.59	0.64	0.60	0.97
HGBR|x0|0.11-159-14	0.325%	0.94	0.59	0.64	0.60	0.97
HGBR|x2|0.07-181-13	0.325%	0.94	0.57	0.65	0.59	0.97
HGBR|x0|0.08-115-13	0.326%	0.94	0.58	0.64	0.60	0.97
HGBR|x2|0.11-159-14	0.328%	0.94	0.58	0.64	0.60	0.97
HGBR|x2|0.12-122-18	0.328%	0.94	0.56	0.66	0.59	0.96
HGBR|x2|0.02-100-16	0.331%	0.95	0.69	0.47	0.56	0.98
HGBR|x0|0.1-102-20	0.331%	0.94	0.55	0.65	0.58	0.96
HGBR|x0|0.07-181-13	0.332%	0.94	0.59	0.64	0.61	0.97
HGBR|x0|0.12-122-18	0.335%	0.94	0.56	0.65	0.59	0.96
HGBR|x0|0.02-100-16	0.337%	0.95	0.70	0.47	0.56	0.99
HGBR|x1|0.05-165-6	0.346%	0.94	0.55	0.59	0.56	0.97
HGBR|x1|0.04-155-6	0.347%	0.94	0.56	0.57	0.56	0.97
HGBR|x1|0.14-181-17	0.353%	0.93	0.50	0.64	0.55	0.96

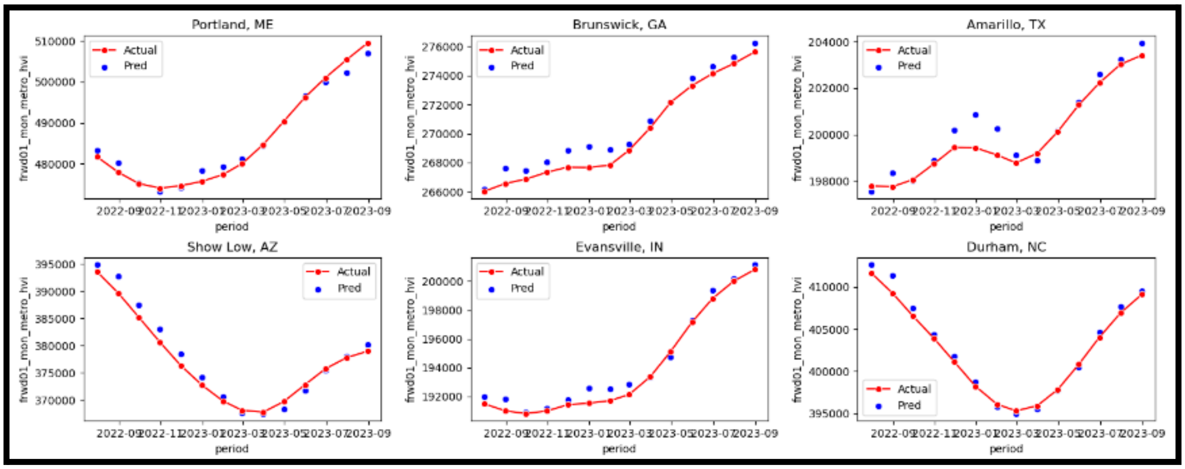
Cross validation results (continued)						
model	mae	accuracy	precision	recall	F1	specificity
HGBR|x1|0.02-100-16	0.358%	0.95	0.63	0.50	0.56	0.98
HGBR|x1|0.04-134-11	0.358%	0.93	0.54	0.61	0.56	0.96
HGBR|x1|0.07-181-13	0.360%	0.93	0.49	0.63	0.53	0.95
HGBR|x1|0.08-115-13	0.364%	0.93	0.48	0.64	0.53	0.95
HGBR|x1|0.1-102-20	0.368%	0.93	0.48	0.66	0.54	0.95
HGBR|x1|0.11-159-14	0.371%	0.93	0.48	0.65	0.54	0.95
HGBR|x1|0.12-122-18	0.373%	0.93	0.49	0.65	0.54	0.95
HGBR|x0|0.01-170-20	0.381%	0.94	0.85	0.27	0.40	1.00
HGBR|x2|0.01-170-20	0.381%	0.94	0.85	0.27	0.40	1.00
HGBR|x0|0.01-120-10	0.387%	0.94	0.84	0.26	0.39	1.00
HGBR|x2|0.01-120-10	0.387%	0.94	0.84	0.26	0.39	1.00
HGBR|x1|0.01-170-20	0.393%	0.94	0.82	0.25	0.38	1.00
HGBR|x1|0.01-120-10	0.398%	0.94	0.82	0.24	0.36	1.00
RF|x2|55-27-sqrt	0.443%	0.96	0.08	0.36	0.17	0.96
RF|x2|77-20-log2	0.447%	0.96	0.02	0.29	0.12	0.96
RF|x2|69-68-sqrt	0.459%	0.96	0.01	0.33	0.03	0.96
RF|x2|12-27-sqrt	0.461%	0.96	0.05	0.29	0.23	0.96
RF|x0|55-27-sqrt	0.469%	0.96	0.05	0.26	0.23	0.96
RF|x0|12-27-sqrt	0.470%	0.96	0.06	0.32	0.15	0.96
RF|x2|28-88-sqrt	0.477%	0.96	-	-	NaN	0.96
RF|x0|69-68-sqrt	0.480%	0.96	0.01	0.32	0.05	0.96
RF|x2|35-67-log2	0.488%	0.96	-	NaN	NaN	0.96
RF|x1|77-20-log2	0.492%	0.96	0.03	0.31	0.07	0.96
RF|x0|28-88-sqrt	0.494%	0.96	-	-	NaN	0.96
RF|x2|71-85-log2	0.496%	0.96	-	NaN	NaN	0.96
RF|x0|77-20-log2	0.499%	0.96	0.01	0.22	0.08	0.96
RF|x2|17-47-log2	0.507%	0.96	0.05	0.82	0.12	0.96
RF|x0|71-85-log2	0.509%	0.96	-	NaN	NaN	0.96
RF|x1|55-27-sqrt	0.512%	0.96	0.03	0.35	0.06	0.96
RF|x2|12-97-log2	0.512%	0.96	-	NaN	NaN	0.96
RF|x1|69-68-sqrt	0.513%	0.96	-	NaN	NaN	0.96
RF|x1|12-27-sqrt	0.515%	0.96	0.03	0.29	0.07	0.96
RF|x0|35-67-log2	0.516%	0.96	-	NaN	NaN	0.96
RF|x1|35-67-log2	0.519%	0.96	-	NaN	NaN	0.96
RF|x1|71-85-log2	0.522%	0.96	-	NaN	NaN	0.96
RF|x1|12-97-log2	0.524%	0.96	-	NaN	NaN	0.96
RF|x0|17-47-log2	0.525%	0.96	0.00	0.13	0.01	0.96
RF|x1|28-88-sqrt	0.530%	0.96	-	NaN	NaN	0.96
RF|x1|17-47-log2	0.532%	0.96	-	-	NaN	0.96
RF|x0|12-97-log2	0.543%	0.96	-	NaN	NaN	0.96
baseline_mean	0.620%	0.92	0.06	0.30	0.10	0.93

The winning model was retrained with the validation training set. The validation test predicted the percentage of home value change within 41% of one standard deviation of the actual home value change for the test period. Out of the 1,827 observations where home value declined, the model correctly predicted a decline 94.3% of the time (recall). Out of the 2,610 observations where the model predicted a decline, only 66% of the time did a decline actually occur (precision). Out of the 4,763 observations where home value increased, the model correctly predicted an increase 81.4% of the time (specificity).

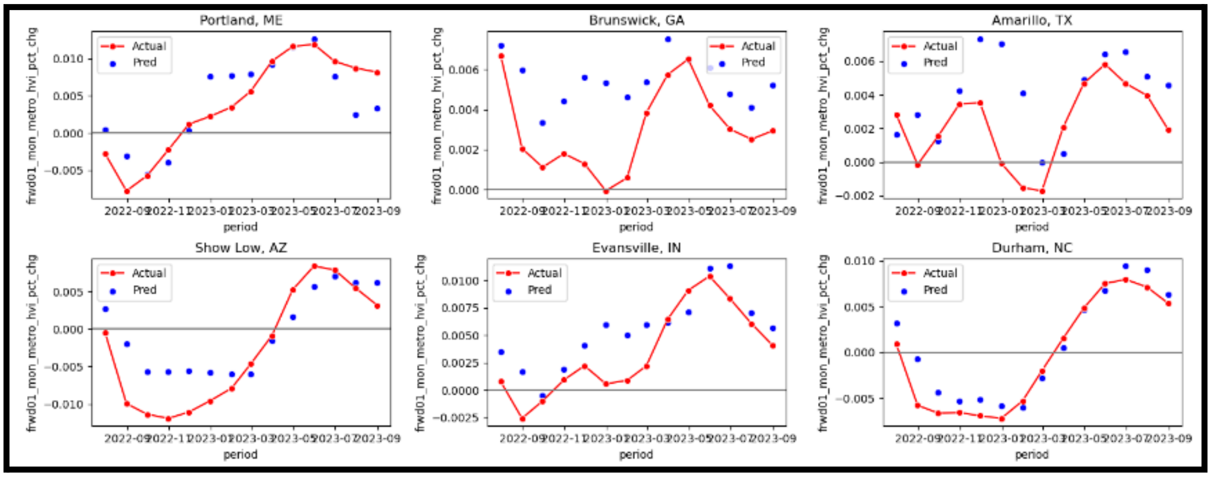


Validation results										
dataset	mae	accuracy	precision	recall	F1	specificity	tp	fp	fn	tn
test	0.258%	0.85	0.66	0.94	0.78	0.81	1723	887	104	3876
train	0.123%	0.97	0.52	0.83	0.64	0.98	431	401	86	17111

After running the final model, metro areas were selected at random to review the predicted results compared to actual home value. Plots for 6 metros have been shown below.



Further comparison was done to review the predicted percentage change compared to the actual percentage change in home value. Plots for the same 6 metros have been shown below. A gray line is included at 0% to show if predictions are directionally aligned with actual change.



Use cases

Realtor

Realtors could drive traffic to their website by providing access to this free model which would allow anyone in the U.S. homemarket to identify the expected direction of home prices in the next month.

Prospective buyers

Prospective buyers thinking about purchasing a house can use the model’s prediction to determine if they should accelerate or decelerate their planned purchase based on the expected direction of the market. If the market is predicted to decline, they could also share this information with the seller to negotiate a lower purchase price.

Prospective sellers

Prospective sellers thinking about selling their house can use the model’s prediction to determine if they should accelerate or decelerate the marketing of their home based on the expected direction of the market. If the market is predicted to increase, they could choose to take additional time for small projects to maximize the readiness of the house while allowing the market to rise.

Future enhancements

1. Reviewing the prediction period to determine how far into the future a reasonable level of accuracy can be maintained.
2. Adding features from additional data sources such as demographic and housing estimates from the US Census Bureau.
3. Additional feature standardization to be explored.
4. Additional models will be explored to determine if the existing model still provides best performance with additional data sources.