

4.4. Feature Importance. As we have described previously and as illustrate in Figure XXX, our models have access to 72 features on each transacting property. We assessed the extent to which some features were more important than others in two steps.

Mean Probability of a Feature Being Included in a Decision Tree
 Mean Probability Across the Entire Ensemble of Decisions Trees
 For Most Accurate Model in Each Training Month

test month	n	prob	feature name
200512	1	35.9	building_living_square_feet
200601	1	19.4	building_living_square_feet
200602	1	35.8	building_living_square_feet
200603	1	16.1	building_living_square_feet
200604	1	16.1	building_living_square_feet
200605	1	30.8	building_living_square_feet
200606	1	35.9	building_living_square_feet
200607	1	35.9	building_living_square_feet
200608	1	19.4	building_living_square_feet
200609	1	16.1	building_living_square_feet
200610	1	19.4	building_living_square_feet
200611	1	20.0	building_living_square_feet
200612	1	19.4	building_living_square_feet
200701	1	16.1	building_living_square_feet
200702	1	16.1	building_living_square_feet
200703	1	16.1	building_living_square_feet
200704	1	16.5	building_living_square_feet
200705	1	16.1	building_living_square_feet
200706	1	19.4	building_living_square_feet
200707	1	9.8	building_living_square_feet
200708	1	16.1	building_living_square_feet
200709	1	9.8	building_living_square_feet
200710	1	20.0	building_living_square_feet
200711	1	16.1	building_living_square_feet
200712	1	20.0	building_living_square_feet
200801	1	16.1	building_living_square_feet
200802	1	19.4	building_living_square_feet
200803	1	16.1	building_living_square_feet
200804	1	16.1	building_living_square_feet
200805	1	16.5	building_living_square_feet
200806	1	16.1	building_living_square_feet
200807	1	16.1	building_living_square_feet
200808	1	16.1	building_living_square_feet
200809	1	16.1	building_living_square_feet
200810	1	16.1	building_living_square_feet
200811	1	16.1	building_living_square_feet
200812	1	16.1	building_living_square_feet
200901	1	16.1	building_living_square_feet
200902	1	19.4	building_living_square_feet

column legend:

test month -> test month
 n -> rank of feature (1 ==> more frequently included)
 prob -> probability feature appears in a decision tree
 feature name -> name of feature

FIGURE 13. For the best-performing models in each training month, the feature most frequently included in the ensembles' decision trees was `building_living_square_feet`. It was included in at least 15 percent of the decision trees except for a few months during the start of the pricing crisis. During the crisis, it remained the most frequently-included feature.

We first examined the most accurate models in each of the training months. The most accurate models were either gradient boosting models or random forests

models, both of which are based on decision trees. We determined for each of the fitted ensemble models, the fraction of decisions trees that contained each feature.

As Figure 13 shows, for the best-performing models in each training month, the feature most frequently included in the ensembles' decision trees was `building_living_square_feet`. It was included in at least 15 percent of the decision trees except for a few months during the start of the pricing crisis. During the crisis, it remained the most frequently-included feature.

We then examined the average rate of inclusion of each feature in all models in the fitted ensembles across all training periods.

As Figure 14 shows, the most-frequently included features were 2 that described the size of the property, then 3 that describe the wealth of the census tract, and then 11 that described the property. Less important than `has_pool` were all features that described other kinds of nearby properties.

Mean Probability of a Feature Being Included in a Decision Tree
 Across the Entire Ensemble of Decisions Trees
 For Most Accurate Model in Each Training Month

mean prob	feature name
19.09	building_living_square_feet
10.69	lot_square_feet
10.16	census2000_median_household_income
7.21	census2000_fraction_owner_occupied
6.41	census2000_avg_commute
5.78	age2
5.70	age
5.05	age_effective
4.79	age_effective2
2.23	building_rooms
1.97	building_baths
1.73	building_bedrooms
1.54	lot_parking_spaces
1.42	building_fireplace_number
0.84	building_stories
0.77	has_pool
0.58	census_tract_has_hotel
0.55	census_tract_has_not_available
0.54	census_tract_has_medical
0.54	census_tract_has_financial_institution
0.54	zip5_has_utilities
0.52	zip5_has_financial_institution
0.50	zip5_has_agriculture
0.49	zip5_has_not_available
0.47	zip5_has_industrial
0.46	census_tract_has_amusement
0.46	census_tract_has_warehouse
0.45	zip5_has_industrial_heavy
0.44	zip5_has_medical
0.44	census_tract_has_utilities
0.37	building_basement_square_feet
0.37	census_tract_has_agriculture
0.36	census_tract_has_residential_condominium
0.35	census_tract_has_office_building
0.35	zip5_has_hotel
0.33	census_tract_has_exempt
0.33	census_tract_has_parking
0.31	census_tract_has_industrial_heavy
0.31	census_tract_has_industrial_light
0.31	zip5_has_transport
0.30	census_tract_has_service
0.29	zip5_has_warehouse
0.29	census_tract_has_industrial
0.28	census_tract_has_transport
0.28	zip5_has_industrial_light
0.27	census_tract_has_duplex
0.27	census_tract_has_any_industrial
0.24	zip5_has_parking
0.24	census_tract_has_vacant
0.23	zip5_has_any_industrial
0.23	census_tract_has_retail
0.20	census_tract_has_any_commercial
0.20	census_tract_has_apartment
0.19	census_tract_has_commercial
0.17	building_is_new_construction
0.15	zip5_has_amusement
0.14	zip5_has_duplex
0.12	census_tract_has_any_non_residential
0.05	zip5_has_apartment
0.03	zip5_has_office_building
0.03	zip5_has_retail
0.02	zip5_has_commercial
0.02	zip5_has_any_commercial
0.01	zip5_has_exempt
0.00	zip5_has_vacant
0.00	zip5_has_residential_condominium
0.00	zip5_has_any_non_residential
0.00	census_tract_has_single_family_residence
0.00	zip5_has_service
0.00	zip5_has_commercial_condominium
0.00	zip5_has_single_family_residence
0.00	census_tract_has_commercial_condominium

column legend:

mean prob -> mean probability feature appears in a decision tree
 feature name -> name of feature

FIGURE 14. The most-frequently included features were 2 that described the size of the property, then 3 that describe the wealth of the census tract, and then 11 that described the property. Less important than `has_pool` were all features that described other kinds of nearby properties.