Is a site inventory predictive of where housing gets built after controlling for covariates? A San Francisco case study.

Salim Damerdji

February 2023

Abstract

Under California housing element law, cities demonstrate they have zoned for enough housing by identifying where housing will get built. With a cross-section dataset of over 150,000 parcels of land in San Francisco for the fourth RHNA cycle, we find that, after controlling for observed confounders, a site's inclusion in the site inventory [is / is not] associated with a higher likelihood of the market building housing on the site.

1 Introduction

Every eight years, each city in California updates its housing element, a planning document whose name stems from being one of seven state-mandated elements of a city's general plan, a long-term planning framework for land-use in a city. In updating a housing element, cities must plan to meet new targets for housing production, viz. the city's Regional Housing Needs Allocation (RHNA). A centerpiece of every housing plan is an inventory of sites where the city shows how it can accommodate enough housing to meet its housing targets.

Historically, site inventories are selected by local planners using local knowledge and heuristics. For smaller towns with low housing targets, planners may hand-select a few vacant or underutilized sites.[1] For larger cities where there are too many parcels to identify underutilized sites one-by-one, planners rely on data-informed heuristics to select sites en-masse. For example, in the RHNA4 cycle (2007-2015), San Francisco selected sites based on the heuristic that a nonvacant site was apt for redevelopment if the existing structure was no larger than 30% of the potential building envelope allowable per the site's zoning.

No matter how a city selects sites, a city's housing element must receive a signoff from the state regulators at the California Department of Housing and Community Development (HCD). Prior to the sixth RHNA cycle, this was

[1]For example, Colma's RHNA5 site inventory had a single site in its inventory, per the Association of Bay Area Government's dataset of housing element site inventories.

largely a paper exercise, as HCD acknowledges in its public-facing presentations. With new state laws like Assemblymember Evan Low's AB 1397 signed into law, HCD became tasked with enforcing side-constraints on how sites could be selected. Nevertheless, these laws were written prior to empirical studies of site inventories and defer to cities on the type of substantial evidence that's appropriate to show nonvacant sites do not contain barriers to redevelopment. While creating new side-constraints - e.g. those related to the size of parcels appropriate for low-income housing - the site inventory selection process is still largely a discretionary activity for cities. One case for this discretion is that cities are better informed as to which sites are likely to be developed and so can use this discretion to select better sites.

Using San Francisco's RHNA4 site inventory as a case study, this paper explores whether the sites selected by planners are associated with a higher rates of housing development after controlling for observed confounders like the parcel's zoning, assessed land value, the age of the property, the neighborhood, and so on. To test this hypothesis, we use both logistic regression and double/debiased machine learning for a logistic partially linear model. With both methods, we will evaluate whether local planners' decisions are associated with a statistically significant difference in outcomes when controlling for observed confounders.

2 Data Sources

This project merges several data sources provided by the City and County of San Francisco.

The BlueSky dataset tracks roughly 150,000 parcels from 2001 to 2016 and includes information on the existing building envelope, the potential buildable envelope given the parcel's zoning designation, historical status, and residential status. Buildings with historical status are harder to build on due to California's Environmental Quality Act; buildings with tenants are harder to build on due to San Francisco's robust tenant protections; and the delta between the existing building envelope and the potential building envelope is correlated with the returns of redevelopment. Thus, this dataset provides several variables that a priori are predictive of where housing will get built. This dataset was, in fact, used by San Francisco in its 6th cycle housing element to identify sites to select for its site inventory.

When the city prepared the BlueSky dataset, they removed duplicative parcel identifiers, including condos; removed parcels with no residential capacity (such as parks); omitted project completions that were not entirely driven by the market; and manually joined parcels to DBI permits. For external researchers without local knowledge, applying these joins between parcels and permits is error-prone and can introduce a larger noise term in the statistical model, which is another reason for relying on San Francisco's BlueSky dataset. It's worth further underscoring that the city removed parcels where the project was not developed due to market forces alone; this alters the interpretation of our findings since, even if inventories are not predictive of where the private market will build housing, inventories may be predictive of where the city intends to fund or otherwise collaborate in building housing.

This dataset is joined with data from the county tax assessor that spans from July 1, 2007 to June 30, 2008. The tax assessor's data includes information on the age of the property, the construction type, the property's square footage, the basement area, lot area, lot shape, the ownership status, the prior sale date of the land, the assessed improvement value, the assessed land value, the number of bedrooms, baths, stories, and units, and more.

Additionally, using San Francisco's Department of Building Inspection's dataset of permits, for each parcel, we know in the preceding eight years how many times the Department of Building Inspection recieved a permit to build, teardown, improve, or alter something on the parcel. A priori, one would think that permits to improve a parcel are negative indicators that the owner is interested in tearing down the property to rebuild. Conversely, it's reasonable that demolition permits are lead indicators for future development on the parcel.

Because steep lots pose unique construction costs, we join this dataset with a topological map of San Francisco. Created in 2019, this dataset is post-treatment but causally unaffected by the treatment, while causally effecting development. Thus, its inclusion increases the precision of our estimate.

An important determinant of economic feasibility is rent, and so we join our dataset with Zillow's Zillow Home Value Index for All Homes (both single family and condos) for January 30, 2007 rent in San Francisco. This is the only neighborhood-level data provided by Zillow.

Because socioeconomic factors affect the political viability of permitting housing in the highly politicized San Francisco housing regime, we also include variables from the 2000 census, which tracks the census tract's age distribution, gender split, racial demographics, homeownership rate, and the density of the tract. We use the year 2000 simply because it is the most recent pre-treatment census date.

Finally, we use the Association of Bay Area Government's dataset on San Francisco's 4th cycle site inventories. These inventories identify where the city claims it can realistically accommodate its housing targets. This is the treatment variable.

In all, our cross-sectional dataset has 153,204 observations with 97 covariates in addition to the treatment and outcome.

3 San Francisco as a Case Study

"The degree to which a parcel is considered built out is measured as its development "softness" and expressed as a percentage of how much of the parcel's potential development capacity is utilized, aggregating residential and non-residential uses. The softness categories in use are 5% and 30%; the categories are mutually exclusive, and a parcel's softness is counted in the category it falls immediately beneath. For example, a parcel that is developed to 20% of its zoned capacity will fall in the 30% softness bracket. "

From page D.7 of housing element: "There were sites which would qualify for a softness label on metrics alone, but for a number of reasons were excluded from the overall softness tally. These cases are listed in Table D-3. These exceptions have been taken largely for practical reasons. For example, fire stations, schools and other public community facilities may be in structures that do not fully utilize the parcels' potential capacity based on underlying zoning standards. These buildings, however, serve a public function and may not likely be turning over for additional development. Similarly, freeways and other dedicated rights-of-way, even if these parcels are zoned for residential uses, are not considered as land suitable for development. Also underutilized parcels that may have residential or mixed uses with at least 10 units are not considered soft for this exercise. It is assumed for the purposes of estimating land inventory that such sites will not likely be demolished and rebuilt. These exemptions, as well as the assumptions and limitations cited in previoius sections, therefore make this a very conservative estimate of the City's remaining capacity." See Table D-3 for more exceptions.

4 Exploratory Data Analysis

Only 0.2% of parcels are developed by the market; and only 2.9% of parcels within the city are planned for growth in SF's fourth RHNA cycle housing element. In other words, both the treatment and outcome are relatively uncommon: most parcels are neither planned for growth nor see growth.

This section includes figures for all variables with a pearson correlation coefficient with an absolute value greater than 0.05.

In Table 1, we see most variables are weakly correlated with the outcome of interest, private residential development. In Table 2, we see most variables are weakly correlated with the treatment as well. A priori, however, one should not expect strong linear, univariate relationships between these variables and development for the simple reason that development is mediated by economic feasibility analyses that involve the non-linear interaction of half a dozen variables, including rent, land values, zoning, and various neighborhood-level factors like impact fees.

Note to self: There must be a better way to format these gargantuan tables.

4.1 Tax Data

A lot of these variables are right-skewed and log transforms are helpful.

Mention outliers, e.g. 555 California Street.

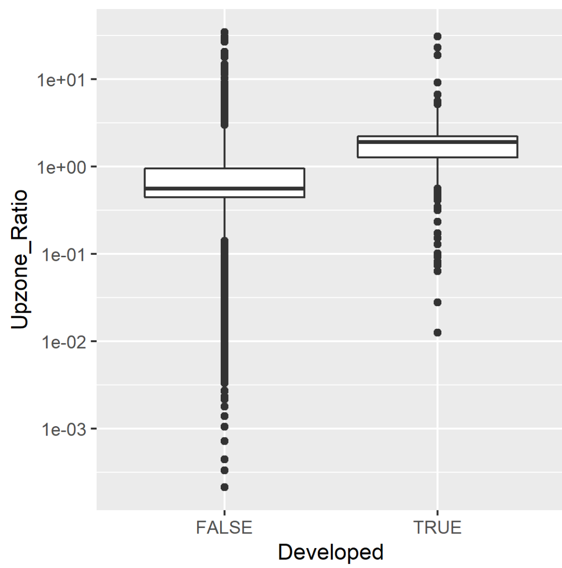

Figure 1: Upzoning ratio (defined as Potential Building Envelope / Existing sq ft) vs development

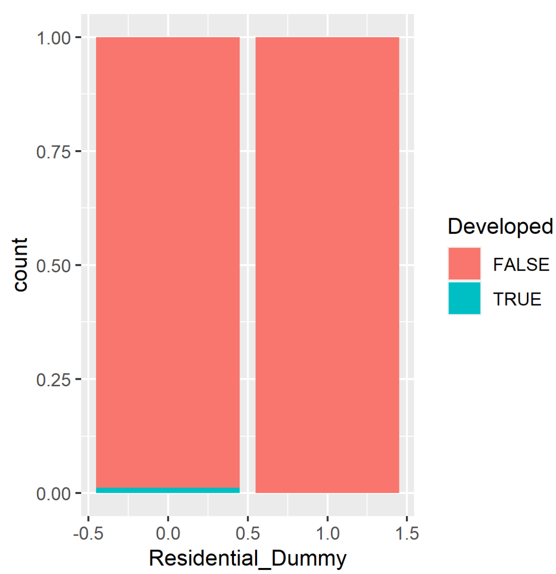

Figure 2: Existing residential use vs development. This bar chart kinda is lacking because so few parcels are developed that it's obscure the relative ratio of development between residential and non-residential. Also, the x axis does not make it clear that this is a binary variable.

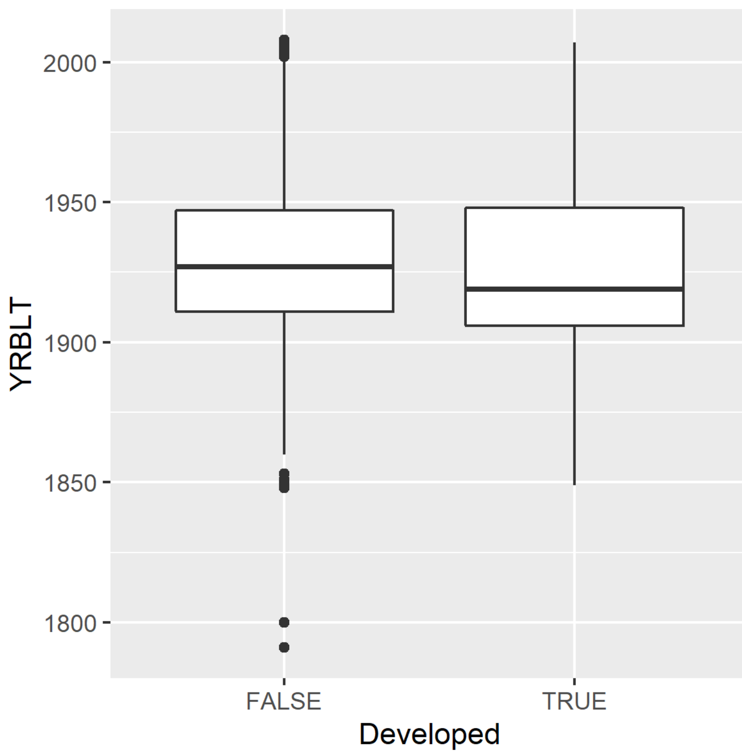

Figure 3: When you omit parcels where the year built is missing (which has a value of 0 in the parcel's file), the existing structure's Year Built and whether it's developed in eight years don't seem strongly related.

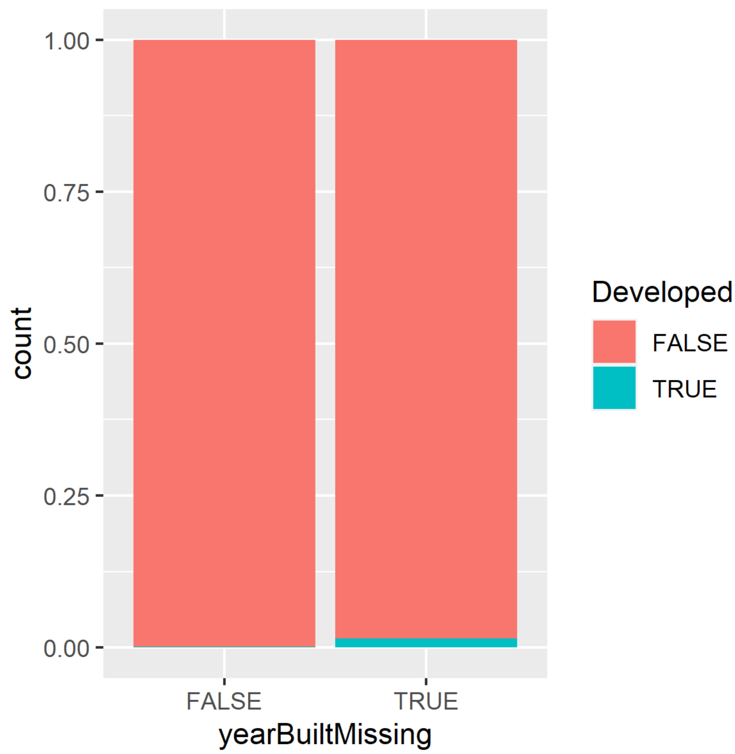

Figure 4: But when the variable YRBLT is missing, it is strongly related with development. This is a case for creating a dummy indicator for missing year built data.

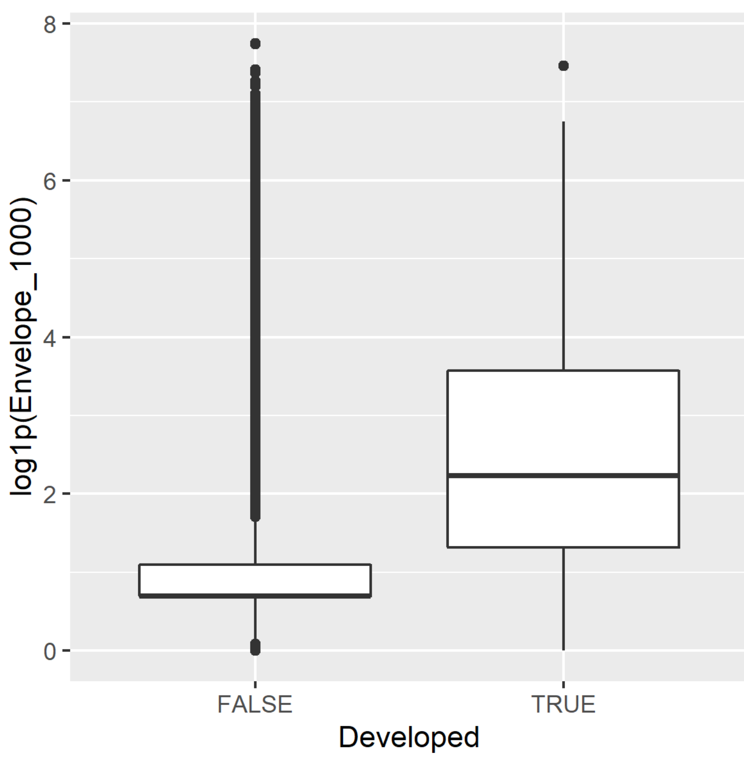

Figure 5: Log of Potential Building Envelope in 1000 sq ft versus development.

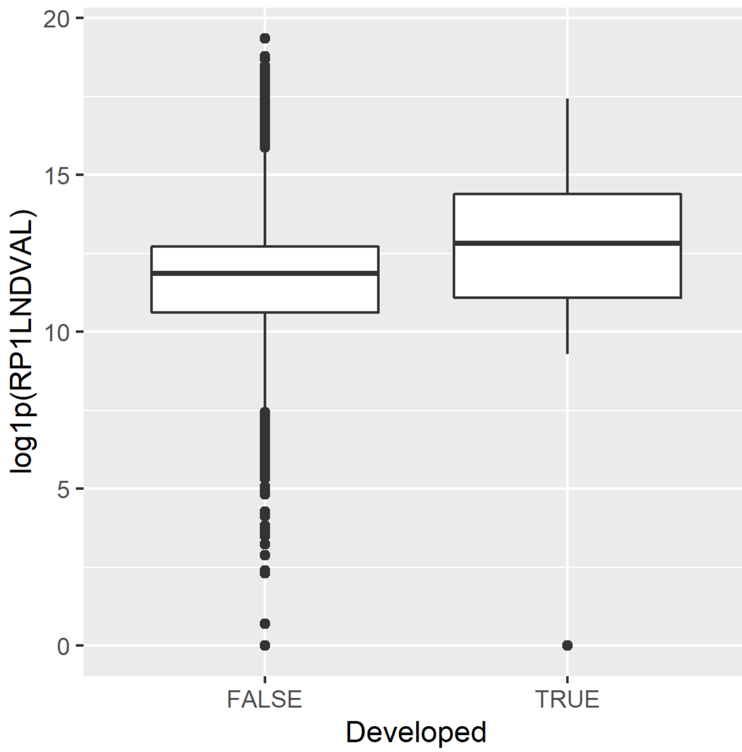

Figure 6: Assessed land value versus development.

Table 1: Pearson correlation with outcome variable, development. Sorted by absolute value of correlation.

Variable	Pearson correlation coefficient
Upzone_Ratio	0.111
Residential_Dummy	-0.076
YRBLT	-0.067
Envelope_1000	0.062
RP1LNDVAL	0.057
zp_DensRestMulti	0.045
owner_occ	-0.042
inInventory	0.039
pipeline6	0.039
zp_Redev	0.034
zp_FormBasedMulti	0.034
renter_occ	0.032
zp_OfficeComm	0.032
pipeline2	0.031
hsehld_1_m	0.029
marhh_chd	-0.027
RP1EXMVL1	-0.026
pipeline1	0.026
zp_PDRInd	0.025
families	-0.025
age_5_17	-0.024
ameri_es	0.024
marhh_no_c	-0.024
BATHS	0.024
ROOMS	0.023
ave_hh_sz	-0.023
age_50_64	-0.022
age_under5	-0.022
vacant	0.021
females	-0.020
pipeline_costs	0.019
age_65_up	-0.019
med_age_f	-0.018
asian	-0.018
age_40_49	-0.016
med_age	-0.016
pop2000	-0.015
ave_fam_sz	-0.014
objectid	0.012
hsehld_1_f	0.012
med_age_m	-0.012
LAREA	0.011
SQFT	0.011
UNITS	0.010
RECURRSALD	-0.010
age_22_29	0.009
rentJan07	-0.009

Table 2: Pearson correlation with treatment variable, inventory site selection. Sorted by absolute value of correlation.

Variable	**Pearson correlation coefficient**
Residential_Dummy	-0.363
YRBLT	-0.243
Upzone_Ratio	0.207
zp_DensRestMulti	0.193
RP1EXMVL1	-0.106
zp_Redev	0.092
ameri_es	0.077
zp_PDRInd	0.074
zp_OfficeComm	0.070
owner_occ	-0.062
rentJan07	-0.056
marhh_no_c	-0.056
Envelope_1000	0.052
zp_RH2	-0.050
pipeline4	0.050
other	0.050
hispanic	0.047
marhh_chd	-0.047
families	-0.046
white	-0.042
females	-0.041
age_50_64	-0.040
Developed	0.039
pipeline2	0.039
STOREYNO	-0.035
pipeline7	0.034
RECURRSALD	-0.034
ROOMS	-0.033
households	-0.031
pipeline1	0.030
hse_units	-0.030
med_age_m	-0.028
pipeline6	0.027
age_65_up	-0.027
pop2000	-0.026
med_age	-0.026
asian	-0.026
zp_FormBasedMulti	0.026
black	0.025
age_40_49	-0.024
zp_RH3_RM1	-0.023
hawn_pi	0.022
age_under5	-0.021
hsehld_1_f	-0.021
BATHS	-0.021
mult_race	0.021
hsehld_1_m	0.021

5 Methods

5.1 Logistic Regression

5.2 Double ML

If planners' selection of sites reflects variables not contained in the dataset, then the treatment - viz., inclusion in the site inventory - is not exogenous given observables, and so a causal interpretation is not warranted. Nevertheless, we can identify the statistical significance of a site's inclusion in the site inventory in a logistic partially linear model, using the gradient boosted classifier CatBoost for both the propensity score model and the outcomes classification model [1][3][2]. After expanding categorical variables with one-hot encoding, the dimensionality of the feature vector is well-over four hundred, without taking into account plausible interactions terms. As a result, using ML to estimate the non-linear portion of this partially linear model reduces variance in estimation given the high-dimensional nuisance parameters.

6 Results

6.1 Baseline Model

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
Intercept	0.00	[0.00, 0.00]	<0.001
inInventory	7.85	[5.64, 10.7]	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

6.2 Logistic Regression Model

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
Intercept	0.00	0.00, 0.00	<0.001
Upzone_Ratio	1.18	1.12, 1.25	<0.001
zp_DensRestMulti	16.3	9.95, 27.4	<0.001
zp_OfficeComm	10.8	5.61, 20.5	<0.001
zp_PDRInd	7.12	3.84, 13.2	<0.001
zp_RH3_RM1	7.48	4.21, 13.2	<0.001
zp_RH2	6.47	3.70, 11.5	<0.001
zp_FormBasedMulti	6.16	2.91, 11.8	<0.001
zp_Redev	6.64	2.81, 14.7	<0.001
pipeline2	1.13	1.00, 1.28	0.033

pipeline3	1.11	1.04, 1.16	<0.001
pipeline4	0.33	0.11, 0.71	0.017
pipeline5	0.58	0.15, 1.05	0.14
pipeline6	1.71	1.26, 2.22	<0.001
pipeline8	0.74	0.65, 0.82	<0.001
log1p(RP1LNDVAL)	1.39	1.29, 1.50	<0.001
log1p(RP1IMPVAL)	0.95	0.91, 0.99	0.012
log1p(ROOMS)	1.10	0.94, 1.26	0.2
log1p(BEDS)	0.85	0.58, 1.16	0.3
DEPTH	0.99	0.98, 1.00	0.007
Historic	0.59	0.42, 0.82	0.002
inInventory	0.84	0.57, 1.21	0.4
Residential_Dummy	0.25	0.17, 0.35	<0.001
Envelope_1000	1.00	1.00, 1.00	0.013
RP1STACDE A	0.00		>0.9
RP1STACDE L	19.5	1.09, 92.4	0.004
RP1STACDE N	21.9	7.34, 65.2	<0.001
RP1STACDE P	0.00	0.00, 0.00	>0.9
RP1STACDE S	0.00	0.00, 0.00	>0.9
log1p(RP1EXMVL1)	0.94	0.89, 0.99	0.028
log1p(FBA)	0.73	0.47, 0.92	0.038
YRBLT	1.00	1.00, 1.00	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

7 Conclusion

8 Appendix

References

[1] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters.* 2018.

[2] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support". In: *CoRR* abs/1810.11363 (2018). arXiv: `1810.11363`. URL: `http://arxiv.org/abs/1810.11363`.

[3] Molei Liu, Yi Zhang, and Doudou Zhou. "Double/debiased machine learning for logistic partially linear model". In: *The Econometrics Journal* 24.3 (2021), pp. 559–588.

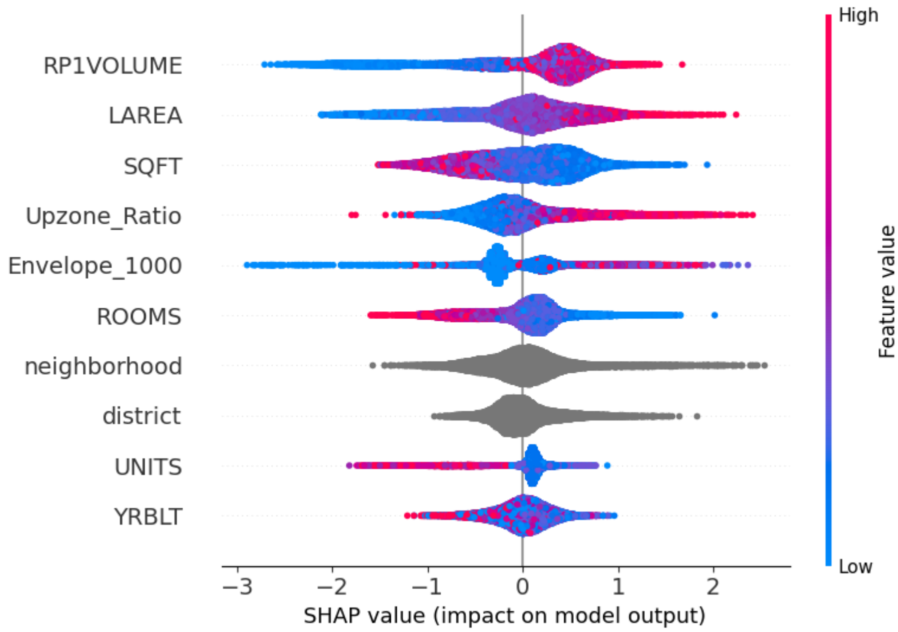

Figure 7: The top 10 most impactful variables on whether a site was included in the site inventory, according to shapley values from the CatBoost propensity score model.

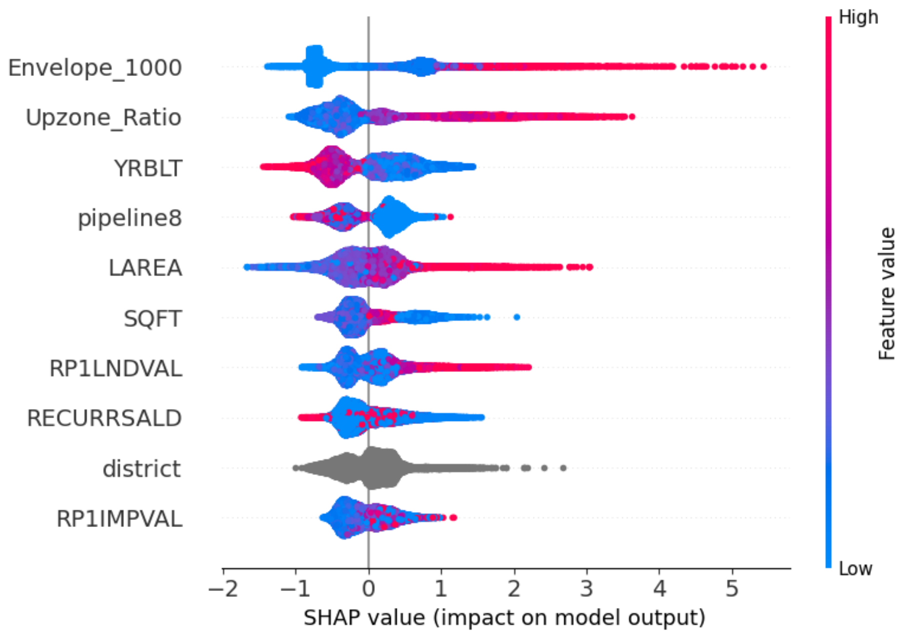

Figure 8: The top 10 most impactful variables on whether a site was developed, according to shapley values from the CatBoost outcomes classification model.

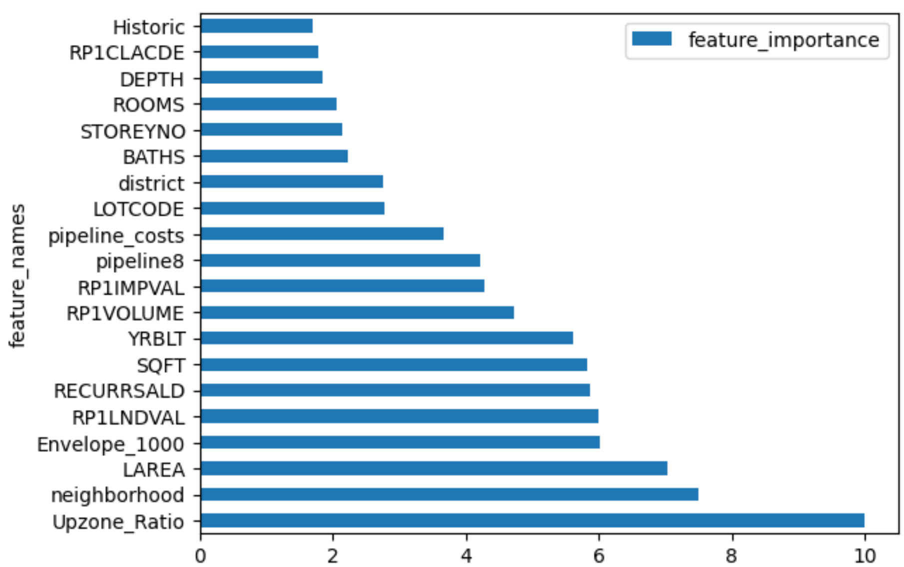

Figure 9: The top 10 most impactful variables on whether a site was developed, according to feature importance values from the CatBoost outcomes classification model.

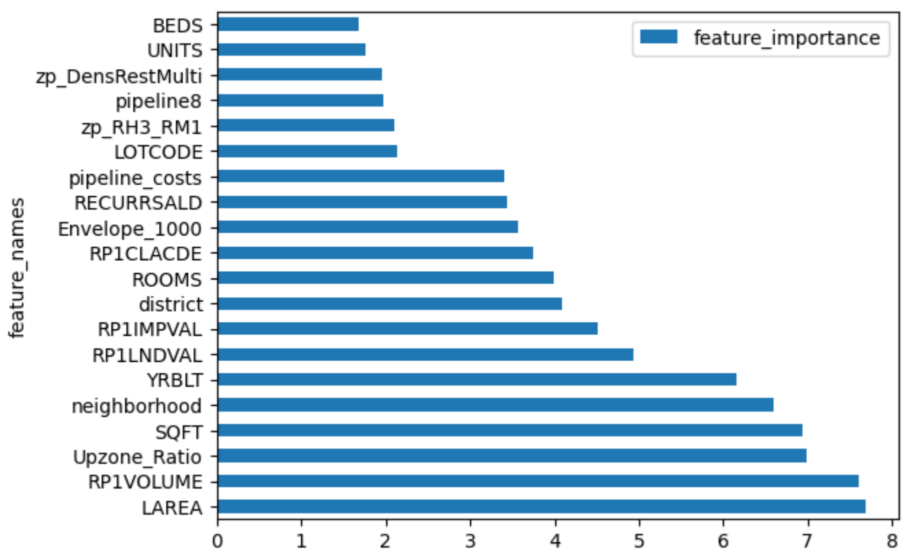

Figure 10: The top 10 most impactful variables on whether a site was selected in the site inventory, according to feature importance values from the CatBoost outcomes classification model.