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 four data sources provided by the City and County of San Francisco with census data and neighborhood-level home price data from Zillow.

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 received 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

San Francisco's fourth RHNA cycle housing element was written in 2009 and adopted on June 29, 2011. While the RHNA cycle began in 2007, a lawsuit related to the Environmental Impact Review for the previous housing element prevented the city from adopting the RHNA4 housing element in a timely way. Work on the site inventory at least extended past April 2009. HCD then approved the draft on April 8, 2011 and re-approved a substantially similar draft on July 29, 2011.

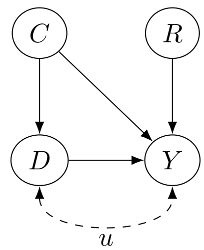

Figure 1: A structural causal model of how selecting a site in the inventory is associated with housing development. Confounders C affect both the treatment D and the outcome Y, while node R causally affects Y alone.

A few words are in order to explain how our data fits into the structural causal model illustrated in Figure 1. In our dataset, every covariate plausibly has a direct causal effect on Y, as we sketched out in Section 2. Many covariates also causally affect the treatment, site selection. For example, the potential buildable envelope on a site is a confounder C since we know from the 2009 Draft that planners looked at the potential buildable envelope divided by the existing buildable envelope to select sites, and it also certainly affects actual development patterns.

Other variables causally affect Y alone and not indirectly through the treatment. For example, we use DBI's permits dataset to track which parcels recently pulled a permit to make an improvement to the existing structure on the land. This surely affects the likelihood of development on the parcel because a recent improvement is indicative that the owner is interested in maintaining the site's existing use. However, because the 2009 Housing Element did not consider these types of permits in selecting sites, we can reasonably surmise that this variable is not a confounder.[2] Nonetheless, we control for variables R because it improves the precision of our estimate.

Notably, all of our covariates that we control for are pre-treatment covariates, so they cannot causally determined by the treatment D. Hence, we avoid introducing bias by controlling for a mediator.

In Figure 1, a causal arrow connects D to Y. However, it's probable that this causal relationship is either nonexistent or negligible. There are at least two ways a site's inclusion in the inventory could be associated with development: 1) planners have their own local knowledge about development patterns and use it to select good sites (or bad sites); 2) a site being included in the inventory causes the site to be more likely to develop courtesy of new state laws. One attractive feature about 2007-2016 is that it was stable for housing element law, so it isolates the question regarding whether planners have better information on development patterns. There was, however, one part of the HAA that coincides with 2007-2016, which was that 20% LI BMR projects were by-right on sites

[2]Note that nothing in our analysis hinges on these variables R *not* being a confounder. If they happen to be confounders, then we control for them all the same.

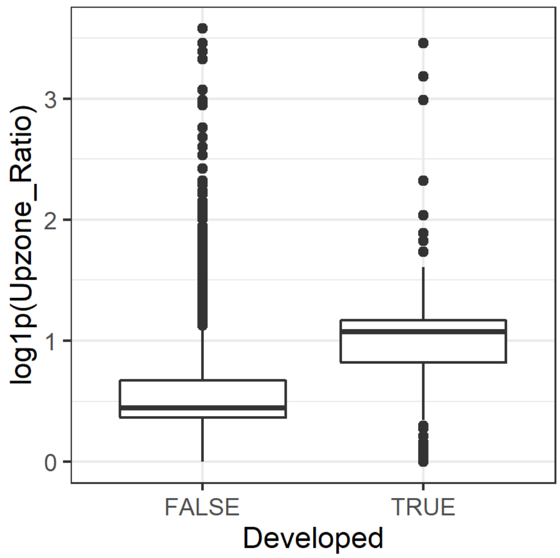

Figure 2: Upzoning ratio (defined as Potential Building Envelope / Existing sq ft) vs development

zoned for LI housing; this was likely not relevant to SF because they probably didn't need to rezone, but we should double-check. If SF did have to rezone, then, if you can observe the rezoning, you'd expect pdev to drop before rezoning and rise afterwards.

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.

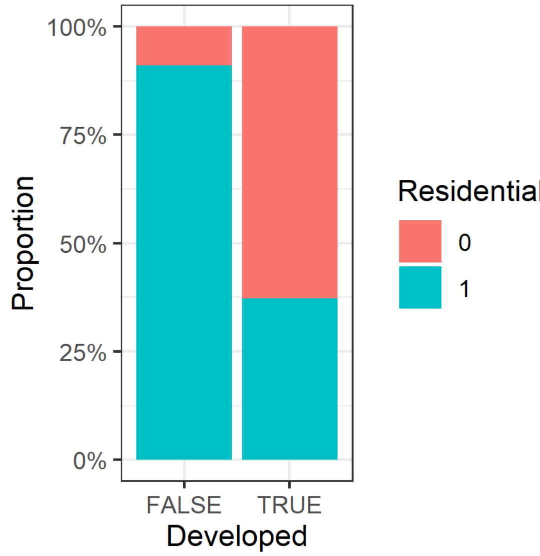

Figure 3: Existing residential use vs development. This bar chart kinda is lacking because so few parcels are developed that it's obscure the relative ratio of development between residential and non-residential. Also, the x axis does not make it clear that this is a binary variable.

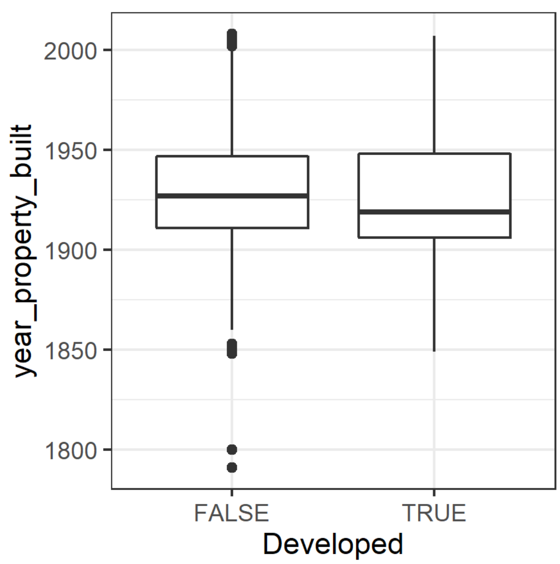

Figure 4: When you omit parcels where the year built is missing (which has a value of 0 in the parcel's file), the existing structure's Year Built and whether it's developed in eight years don't seem strongly related.

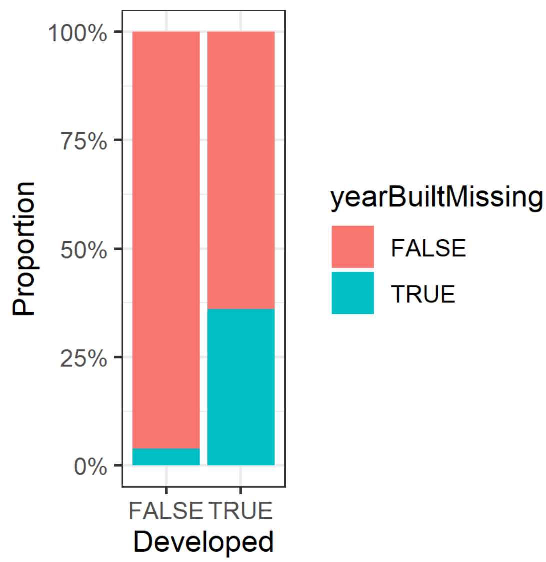

Figure 5: But when the variable YRBLT is missing, it is strongly related with development. This is a case for creating a dummy indicator for missing year built data.

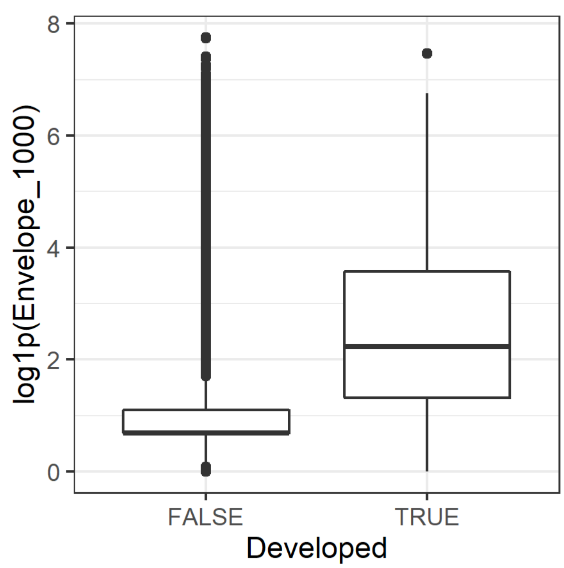

Figure 6: Log of Potential Building Envelope in 1000 sq ft versus development.

Table 1: Pearson correlation with outcome variable, development. Sorted by absolute value of correlation. Only twenty variables with largest abs(corr) are shown.

Variable	**Pearson correlation coefficient**
Upzone_Ratio	0.112
Residential_Dummy	-0.076
year_property_built	-0.067
yearBuiltMissing	0.067
buildable_envelope_1000	0.062
assessed_land_value	0.054
zp_DensRestMulti	0.046
census_vacant	0.044
permit_demolitions	0.039
inInventory	0.038
census_renter_occ	0.038
census_hsehld_1_m	0.037
census_native	0.035
census_females	-0.035
census_marhh_chd	-0.035
zp_Redev	0.034
zp_FormBasedMulti	0.034
census_families	-0.032
permit_new_construction_wood_frame	0.031
zp_OfficeComm	0.030

Table 2: Pearson correlation with treatment variable, inventory site selection. Sorted by absolute value of correlation. Only twenty variables with largest abs(corr) are shown.

Variable	**Pearson correlation coefficient**
Residential_Dummy	-0.363
yearBuiltMissing	0.246
year_property_built	-0.246
Upzone_Ratio	0.207
zp_DensRestMulti	0.193
homeowner_exemption_value	-0.106
census_native	0.094
zp_Redev	0.092
census_females	-0.090
x_coord	0.075
zp_PDRInd	0.074
zp_OfficeComm	0.070
census_families	-0.067
census_marhh_no_c	-0.067
census_vacant	0.065
census_mult_race	0.065
census_hispanic	0.059
census_marhh_chd	-0.059
home_prices_jan07	-0.056
census_age_50_64	-0.053

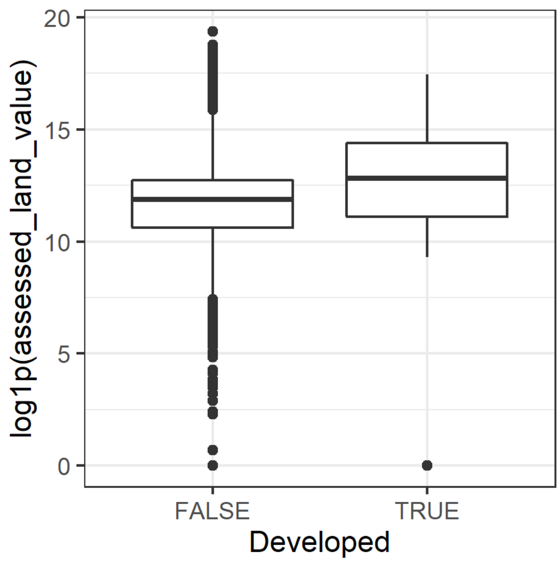

Figure 7: Assessed land value versus development.

5 Methods

5.1 Logistic Regression

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

In diagnostic section, mention outliers, e.g. 555 California Street.

5.2 Double ML

Consider the logistic partially linear model:

$$P(Y_i|D_i, X_i) = expit\{d_i\beta + g(X_i)\} \tag{1}$$

where $Y_i \in \{0, 1\}$ is a binary indicator of whether parcel i was developed, D_i indicates whether parcel i was included in the site inventory, $B \in \mathbb{R}$, X_i is the set of pre-treatment covariates (including observed confounders), and $g(.)$ is a non-linear function.

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

First we consider a model that only looks at whether a site is in the inventory. This is effectively the model the state relies on by relying on site inventories. Without controlling for other variables, the odds of development is associated with somewhere between 5.6x and 10.7x for sites that are selected in the inventory. This is a large, practically significant effect, with a vanishingly small p-value. The model, however, only explains 2.8% of the uncertainty in the dataset, as indicated by the r-squared deviance.

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
(Intercept)	0.00	0.00, 0.00	<0.001
inInventory			
FALSE			
TRUE	7.85	5.64, 10.7	<0.001
		RSq.kl:	.028
		AIC:	3634
		BIC:	3654

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

6.2 Logistic Regression Model

This model explains 27% of the uncertainty in the dataset, indicating this is a far superior model for understanding development patterns than relying on the site inventory alone. Many variables have odds-ratios with point estimates greater than the point estimate for inInventory. Reduces both AIC and BIC compared to baseline model.

The treatment variable inInventory is associated with a 16% reduction in the odds of development of a parcel. This is hard to explain. The sign of the coefficient is the opposite direction as the baseline model.

Note that this model was selected before I added census data, slopes data, and rent data. It may well no longer hold. Furthermore, this model was selected in a somewhat haphazard way. I used forward stepwise selection with respect to AIC, introduced log transforms after the fact, and dropped two categorical variables that create instability in fitting the data. Given each step of the forward stepwise selection chose between dozens of variables to add, multiple testing is a concern, and I am also somewhat skeptical of the inference after model selection, which should increase the standard errors. Finally, I am skeptical of the model because it lacks interaction terms to approximate the sort of considerations in a pro forma, which we know should be a part of any plausible model. I am including the model only as a placeholder.

Variable	**Odds Ratio**	**95% Confidence Interval**	**p-value**
Intercept	0.00	0.00, 0.00	<0.001
inInventory	0.84	0.57, 1.21	0.4
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
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
		RSq.kl:	0.267
		AIC:	2780
		BIC:	3118

6.3 Double/Debiased ML

7 Conclusion

8 Appendix

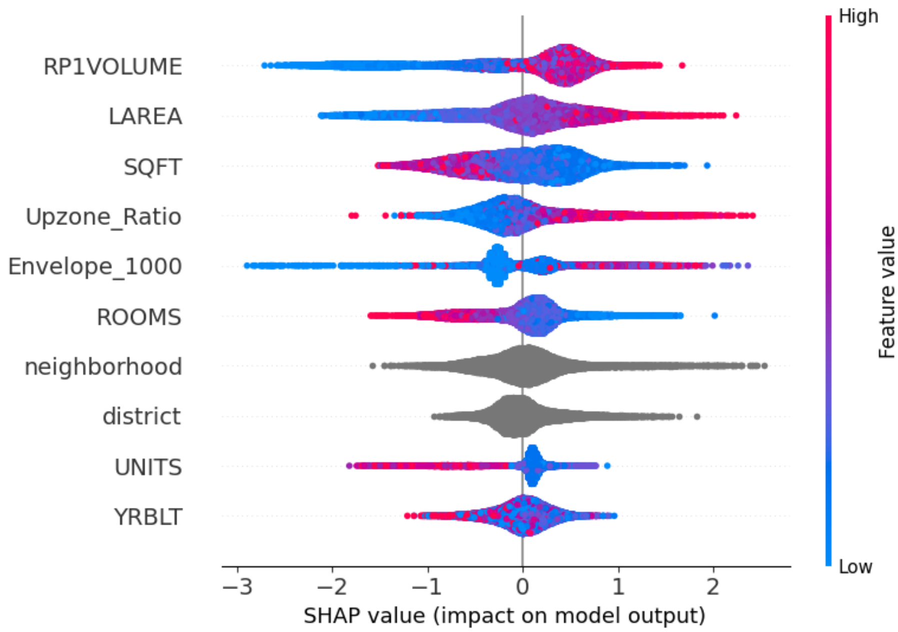

Figure 8: The top 10 most impactful variables on whether a site was included in the site inventory, according to shapley values from the CatBoost propensity score model.

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.

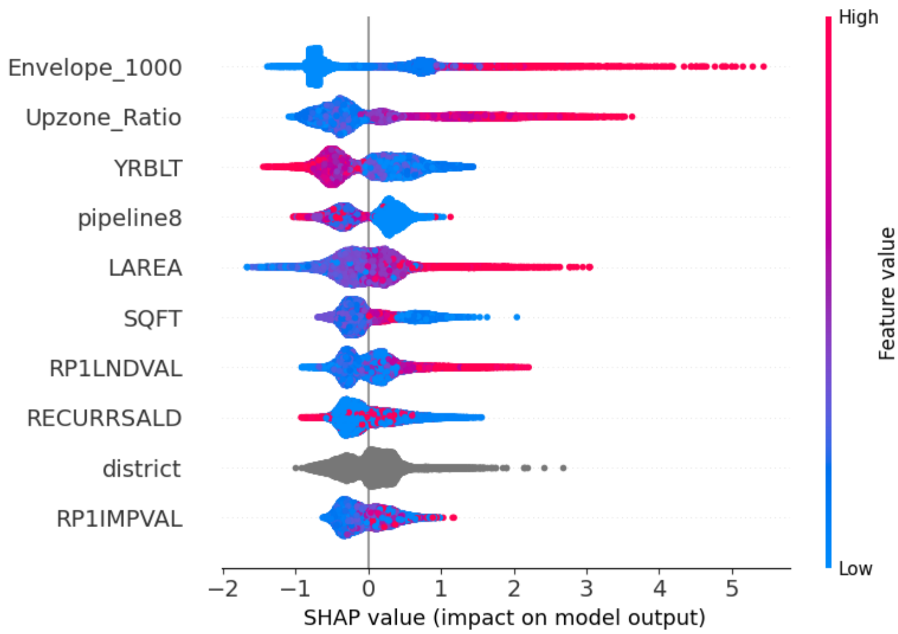

Figure 9: The top 10 most impactful variables on whether a site was developed, according to shapley values from the CatBoost outcomes classification model.

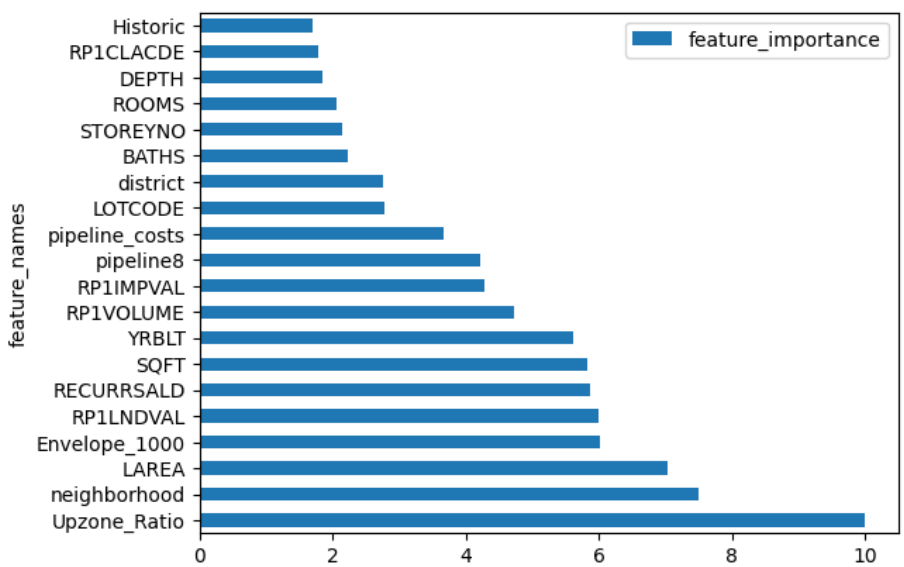

Figure 10: The top 10 most impactful variables on whether a site was developed, according to feature importance values from the CatBoost outcomes classification model.

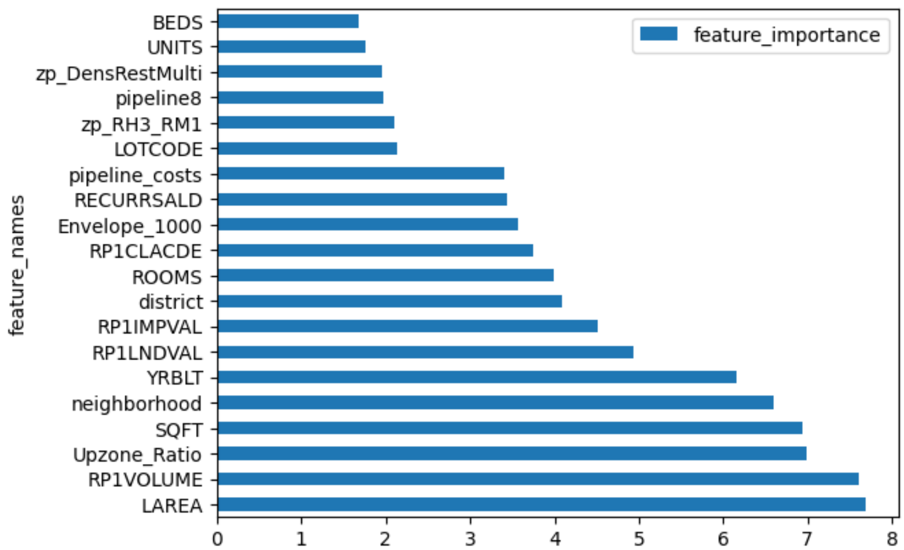

Figure 11: The top 10 most impactful variables on whether a site was selected in the site inventory, according to feature importance values from the CatBoost outcomes classification model.