When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions

by

Kristen M. Altenburger and Daniel E. Ho

Online Appendix

A.1 311 Call Complaint Data in New York

New York Data Processing. In order to merge the inspection and complaint data sets, we process the New York data in several steps.

First, we standardize address formats across the two data sets. We use the "incident address" in the complaint data and the establishment address in the inspection data. We standardize addresses by using consistent street abbreviations (e.g., "St" for "Street", "Rd" for "Road"), cardinal directions (e.g.,"N" for "North"), and street numbers (e.g., "1" for "1ST" or "First").

Second, we subset addresses in the inspection data that resolve to unique establishments. This comprises roughly 84% of all establishments in the inspection data. Examples of addresses that are not unique to an establishment include addresses of food courts and airports.

Third, because 17% of 311 calls are missing address information, these units are ineligible to be matched to the inspection data.

We then match based on exact addresses. To account for temporal sequencing, we match each complaint to the proximate prior routine inspection. This process results in 44% of complaints being matched to a unique establishment in the inspection data.[1] In order to validate this process, we drew random samples of unmatched addresses. One of the principal reasons for why a large set of complaints cannot be matched to inspection data is that 311 complaints are often issued for establishments that fall under the regulatory jurisdiction of the state department of agriculture, not the city department of health (e.g., delis, meat, convenience stores, supermarkets).

[1] This excludes complaints without any address information.

Table A1
New York Ethnicity Coding

Asian	Count	Asian	Count
Chinese	7281	Chinese/Japanese	140
Japanese	2213	Pakistani	121
Indian	1015	Chinese/Cuban	59
Asian	883	Afghan	40
Thai	866	Indonesian	31
Korean	662	Polynesian	8
Vietnamese/Cambodian/Malaysia	241		

Non-Asian	Count	Non-Asian	Count
American	16272	Soul Food	182
Pizza	3590	Continental	163
Cafe/Coffee/Tea	3485	Barbecue	147
Italian	3372	Salads	135
Latin	3236	Bangladeshi	132
Mexican	2593	German	109
Caribbean	2302	Creole	104
Bakery	2284	Soups & Sandwiches	103
Spanish	1961	Filipino	101
Pizza/Italian	1552	Polish	97
Donuts	1530	Brazilian	93
Chicken	1259	Tapas	88
Hamburgers	1240	Armenian	85
Sandwiches	1060	Ethiopian	53
Jewish/Kosher	1053	Pancakes/Waffles	53
Delicatessen	1040	English	49
French	1001	Hot Dogs	45
Ice Cream, Gelato, Yogurt, Ices	729	Moroccan	44
Juice, Smoothies, Fruit Salads	677	Australian	42
Irish	667	Portuguese	36
Mediterranean	579	Egyptian	32
Sandwiches/Salads/Mixed Buffet	496	Hawaiian	27
Middle Eastern	494	Hot Dogs/Pretzels	21
Bagels/Pretzels	493	Southwestern	19
Seafood	467	Cajun	18
Tex-Mex	436	Fruits/Vegetables	15
Greek	422	Scandinavian	15
Other	302	Not Listed/Not Applicable	13
Russian	267	Creole/Cajun	11
Peruvian	263	Czech	11
Vegetarian	245	Chilean	10
Steak	244	Iranian	9
Eastern European	241	Nuts/Confectionery	9
African	239	Californian	4
Turkish	200	Soups	3
Bottled beverages	189		

A.2 Yelp Review Data in King County from Kang et al. (2013)

Table A2
King County Ethnicity Coding

Asian	Count	Asian	Count	Asian	Count
Chinese	1173	Dim Sum	243	Himalayan/Nepalese	23
Japanese	1010	Korean	108	Cambodian	19
Vietnamese	802	Pakistani	72	Indonesian	16
Thai	725	Hawaiian	69	Mongolian	14
Sushi Bars	602	Filipino	42	Shanghainese	13
Asian Fusion	285	Szechuan	35	Malaysian	11
Indian	275	Taiwanese	29	Laotian	7
Cantonese	250	Hot Pot	25		

Non-Asian	Count	Non-Asian	Count	Non-Asian	Count
Sandwiches	1523	Middle Eastern	97	Moroccan	25
Pizza	1144	Tapas Bars	92	Food Court	21
Mexican	968	Soup	82	African	20
Breakfast & Brunch	881	Tex-Mex	73	Belgian	19
Italian	838	Southern	70	Brazilian	19
Fast Food	788	Modern European	68	Fondue	13
Seafood	700	Hot Dogs	67	Persian/Iranian	12
Burgers	572	Caribbean	66	Afghan	11
Delis	566	Soul Food	64	Haitian	11
Cafes	533	Latin American	55	Trinidadian	11
Mediterranean	445	Spanish	55	Kosher	10
Barbecue	293	Chicken Wings	53	Polish	10
Greek	267	Turkish	49	Salad	9
Vegetarian	224	Food Stands	47	Lebanese	7
French	211	Basque	46	Scandinavian	7
Buffets	181	Cheesesteaks	42	Comfort Food	4
Ethiopian	168	Gastropubs	41	Puerto Rican	4
Diners	144	British	41	Venezuelan	3
Vegan	141	Tapas/Small Plates	40	Colombian	3
Irish	135	Cajun/Creole	38	Australian	2
Creperies	133	Live/Raw Food	36	Scottish	2
Gluten-Free	132	Salvadoran	33	Senegalese	2
Steakhouses	129	Cuban	28	Egyptian	1
German	105	Russian	26		
Fish & Chips	98	Halal	26		

A.2.1 King County Robustness Checks

Table A3 provides similar results from a broader classification of "ethnic" (i.e., non-European) cuisines. As before, the coefficients are consistently positive for ethnic establishments, predicting a higher (conditional) probability of a suspicious search term.

Table A3
Logistic Regression Estimates with Explanatory Variable for "Ethnic" Cuisine, instead of Asian Cuisine

	Model 1	Model 2	Model 3	Model 4	Model 5
Ethnic	0.26***	0.37***	0.32***	0.33***	0.32***
	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)
Prior score		0.00			
		(0.00)			
Avg. prior score			0.01***	0.01***	0.01***
			(0.00)	(0.00)	(0.00)
Review count		0.03***	0.03***	0.03***	0.03***
		(0.00)	(0.00)	(0.00)	(0.00)
Review rating				–0.47***	–0.46***
				(0.04)	(0.04)
Year FE	no	yes	yes	yes	yes
ZIP FE	no	no	no	no	yes

Notes: Similar to earlier results, conditional on the same inspection violation score, ethnic establishments receive higher violation scores.

Figure A1 assesses to what extent effects might be driven by particular ZIP codes. Given that inspectors are principally assigned by ZIP code, one might worry whether lenient inspectors are disproportionately assigned to ZIP codes with many Asian establishments. These results fit logistic regressions to all large ZIP codes with sufficient numbers of Asian and non-Asian establishments, showing that the average effect is not driven by any particular ZIP code.

A.2.2 Counts of Suspicious Search Terms

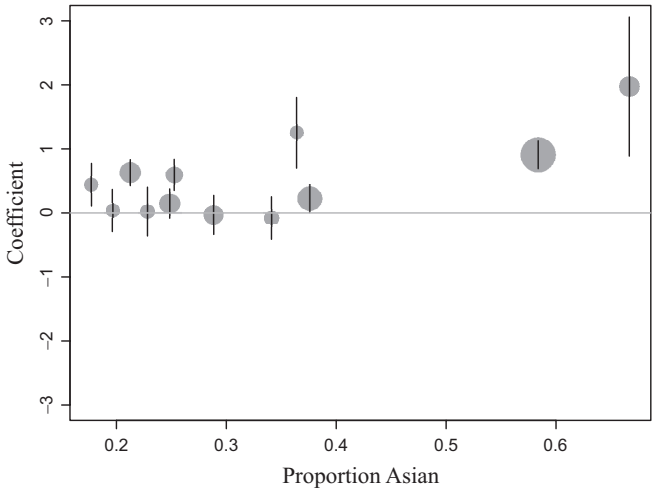
The terms raising suspicion of foodborne illness by New York are: sick, vomit, diarrhea, and food poisoning, presented in Table A4.

Table A4
Counts of Suspicious Search Terms in King County Data

	"food poisoning"	"vomit"	"diarrhea"	"sick"
Asian	55	18	10	344
Non-Asian	98	85	17	539
Prop.	0.36	0.17	0.37	0.39

Notes: Terms were identical to the ones used by New York City in initiating investigations based on Yelp reviews. Prop. indicates the proportion of reviews that are for Asian establishments in all appearances of the term, when the baseline percentage of Asian establishments is 28%.

Figure A1
Effect by ZIP Code



Notes: This figure displays coefficient estimates for each ZIP code, plus or minus standard error, against the proportion of inspections conducted for Asian establishments. These estimates come from separate logistic models fitted to each ZIP code with at least 100 Asian and 100 non-Asian establishments, controlling for the average prior inspection score and year fixed effects. These results demonstrate that the difference in the probability of a suspicious Yelp term is not driven by any single ZIP code and that it is not driven by the proportion of Asian establishments.

A.3 Quantile–Quantile Plot Comparing Asian and Non-Asian Establishments in King County and New York

Figure A2 shows the inspection performance difference is substantial.[2]

A.4 Application of Pope and Sydnor (2011)

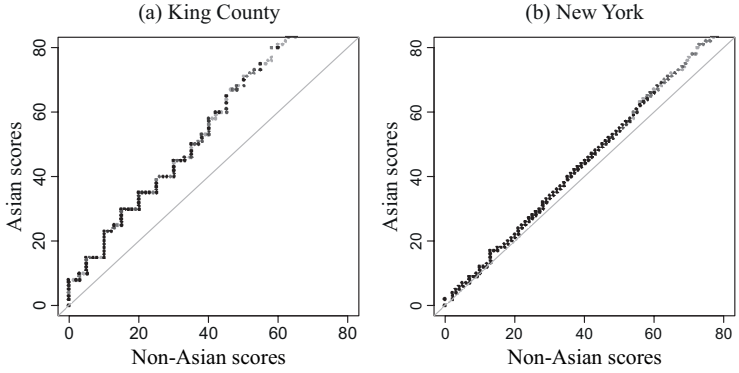
A.4.1 Hyperparameter Tuning

This section describes how we tune the random forest models via cross-validation that is implemented for both the Monte Carlo simulations and analysis for New York and Seattle. We tune the following hyperparameters[3] for RF when applied to NY or WA, and for the simulation piece, we limit `n_estimators = 100`, `max_depth = [5,10]`, and `min_samples_split = [2,5,10]` to help with run time:

[2] It is worth noting that the scales are not directly comparable across jurisdictions, as violations and scoring differ across jurisdictions.

[3] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, accessed September 21, 2018.

Figure A2
Quantile–Quantile Plot Comparing Violation Score Distribution across Asian and Non-Asian Establishments



Notes: This performance difference makes it important to control for inspection score histories to assess for bias in Yelp reviews. For visibility (as a small number of outliers exist), we truncate the axes, but roughly 99.9% of New York's inspections and 99.6% of King County's score below 80 points.

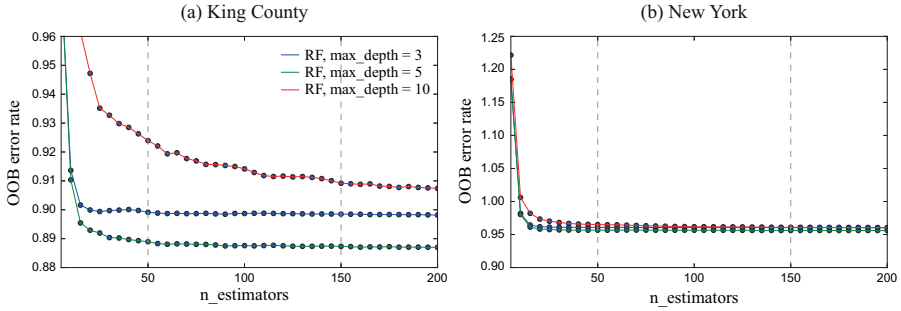
```
max_depth = [3,5,10]
max_depth.append(None)
min_samples_leaf = [5,10,20,50,100]
min_samples_split = [2,3,4,5,10]
n_estimators = [50,100,150]
max_features = ['auto',0.25,0.5,0.75]

random_grid = {'max_depth': max_depth,
               'min_samples_leaf': min_samples_leaf,
               'max_features': max_features,
               'n_estimators': n_estimators,
               'min_samples_split': min_samples_split}
```

The parameters that are typically noted to be more likely to affect overfitting are `n_estimators` (more DT tends to decrease the chance of overfitting), `max_features` (using fewer features to split helps with overfitting), `max_depth` (determines how far the trees are allowed to grow), and `min_samples_leaf` (can result in overfitting if too small).

Then with this search space of parameters, we implement a cross-validated grid-search approach to search exhaustively over the training data set and then to select hyperparameters based on this search. Note that Python's implementation of grid search has the `cv` parameter,[4] which allows the training data to be internally split for validat-

[4] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, accessed September 21, 2018.

Figure A3
Comparison of Hyperparameter Search Range (`n_estimators` and `max_depth`) on Predictive Performance



Notes: For King County on one train split, we visually illustrate the range of `n_estimators` we consider as marked by the region between the two vertical gray lines. We see in this range that the OOB error rate has started to level off. We provide an analogous visual for New York.

ing the model parameters. In the main paper, we report final results on the test data set, given the selected hyperparameters.

A.4.2 Additional Monte Carlo Simulation: Class Imbalance and SAP–SUP Correlation

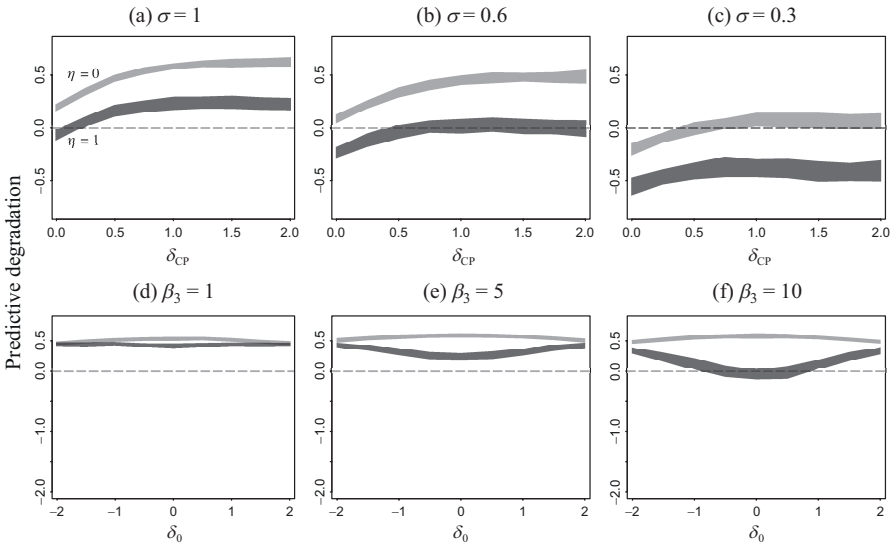
Here we provide additional results from the Monte Carlo simulation. As a reminder from the setup in the main paper, we assume the following data-generating process (DGP), where we introduce the parameters in boxes:

$$
\begin{aligned}
X_i^{\text{CP}} &\sim N(0,1);\\
X_i^{*\text{SUP}} &= \frac{1}{1+\exp[-(\delta_0+\delta_{\text{CP}}X_i^{\text{CP}}+v_i)]};\\
X_i^{\text{SUP}} &= \begin{cases} 1 & \text{if } X_i^{*\text{SUP}} > 0.5,\\ 0 & \text{otherwise;}\end{cases}\\
X_i^{\text{SAP}} &\sim N\big(\boxed{\eta}\times X_i^{\text{SUP}}, \boxed{\sigma}\times X_i^{\text{SUP}} + \big(1-X_i^{\text{SUP}}\big)\big);\\
y_i &= \beta_0+\beta_1 X_i^{\text{SAP}}+\beta_2 X_i^{\text{CP}}+\beta_3 X_i^{\text{SUP}}+\boxed{\beta_4}\big[X_i^{\text{SUP}}\times\big(X_i^{\text{SAP}}+2\big)^2\big]+\varepsilon_i.
\end{aligned}
$$

The top row of Figure A4 shows that as the correlation between SUP and CP increases ($\delta_{\text{CP}} > 0$), the performance gap between the restricted and proposed approaches decreases. This makes sense because SUP imbalance will be most acute when CP predictors are orthogonal to SUP.[5] The bottom row of Figure A4 examines class imbalance along SUP lines (i.e., when the proportions of units with SUP = 1 and 0

[5] Here all other parameters are fixed at $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 5$.

Figure A4
Effect of SAP–SUP Correlation (δ_{CP}) and SUP–Class Imbalance (δ_0)



Notes: In the subfigures (a) through (c), we observe that as δ_{CP} increases that marginalization is more robust to SAP nonoverlap. Subfigures (d) through (f): As we vary the relative class proportions of each SUP class, we observe that the setting where SUP classes are balanced results in the largest performance gap between restricted vs. P&S approaches.

are not equal).[6] Class imbalance, when $|\delta_0| > 0$, generally improves the accuracy of marginalization.[7] The intuition behind this result is that extrapolation becomes problematic for fewer observations than in the balanced class setting.

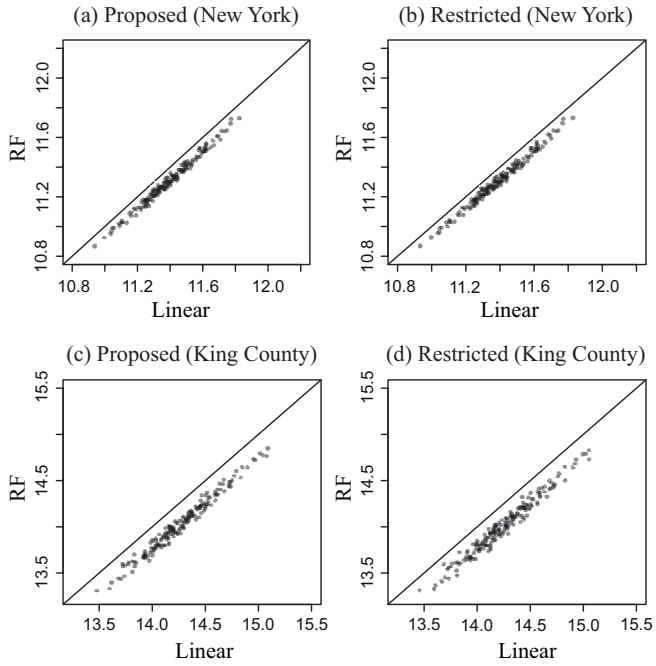
A.4.3 Comparison between Linear Regression and Random Forest Models

In Figure A5, we illustrate on both New York and King County the RMSE of a random forest model versus linear regression for both the proposed and restricted approaches across 200 iterations of 80–20 train–test splits. We see that in both cases, the random forest has better predictive performance than the linear regression model.

[6] Here all other parameters are fixed at $\beta_1 = 1$, $\beta_2 = 1$, and $\eta = 0$.

[7] We note that class imbalance also illustrates the challenge of overall accuracy; class imbalance can hide poor accuracy on the minority class.

Figure A5
RMSE of a Random Forest Model versus Linear Regression

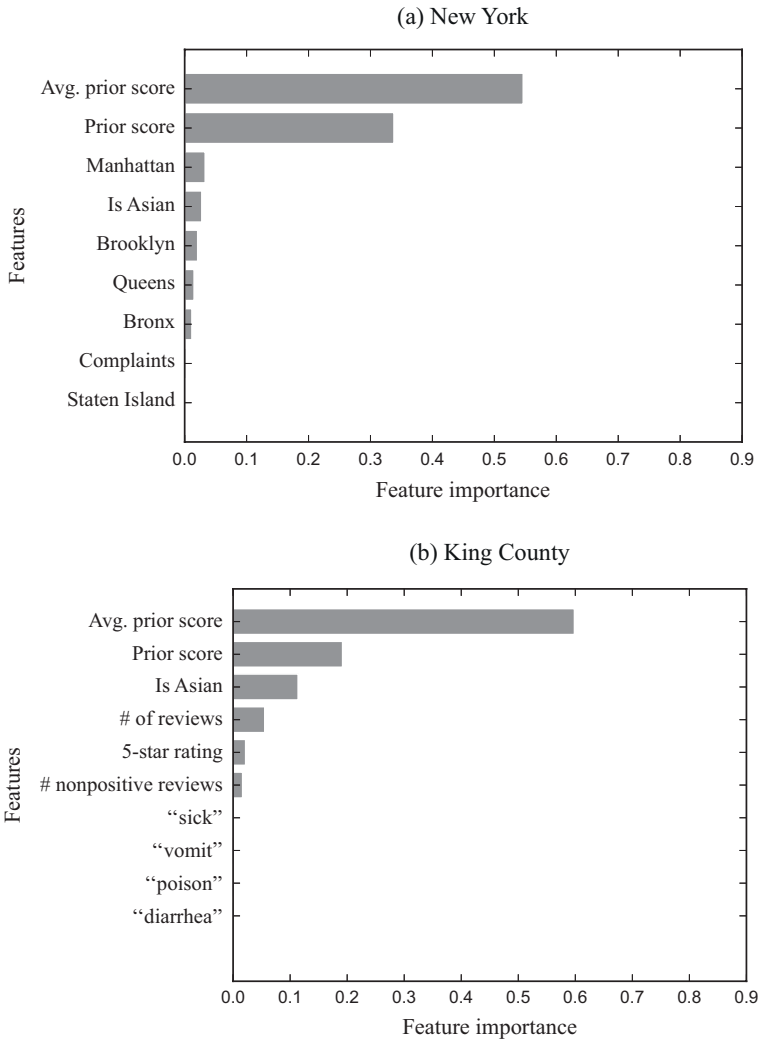


A.4.4 Feature Importance

We calculate feature importance for each application on a single train–test split.[8] As illustrated in Figure A6, we observe that inspection history is a strong predictor.

[8] We acknowledge that there are many different measures of feature importance. We implement the feature-importance calculation score implemented in `scikit-learn`, which is a form of Gini importance that measures the decrease in Gini impurity due to splitting on a feature, and is averaged across all trees in the ensemble. We refer the reader to Louppe et al. (2013) for a more general discussion of feature-importance measures.

Figure A6
Visual of Feature Importance for New York and King County

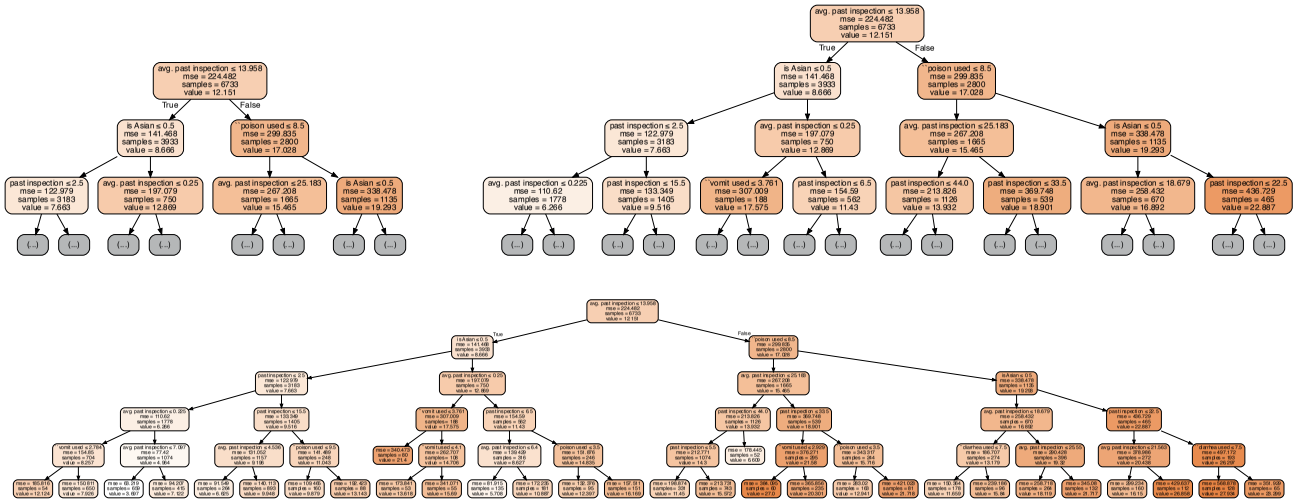


Notes: We examine feature importance for New York and King County on one train–test split.

A.4.5 Sample Trees from Random Forest Models

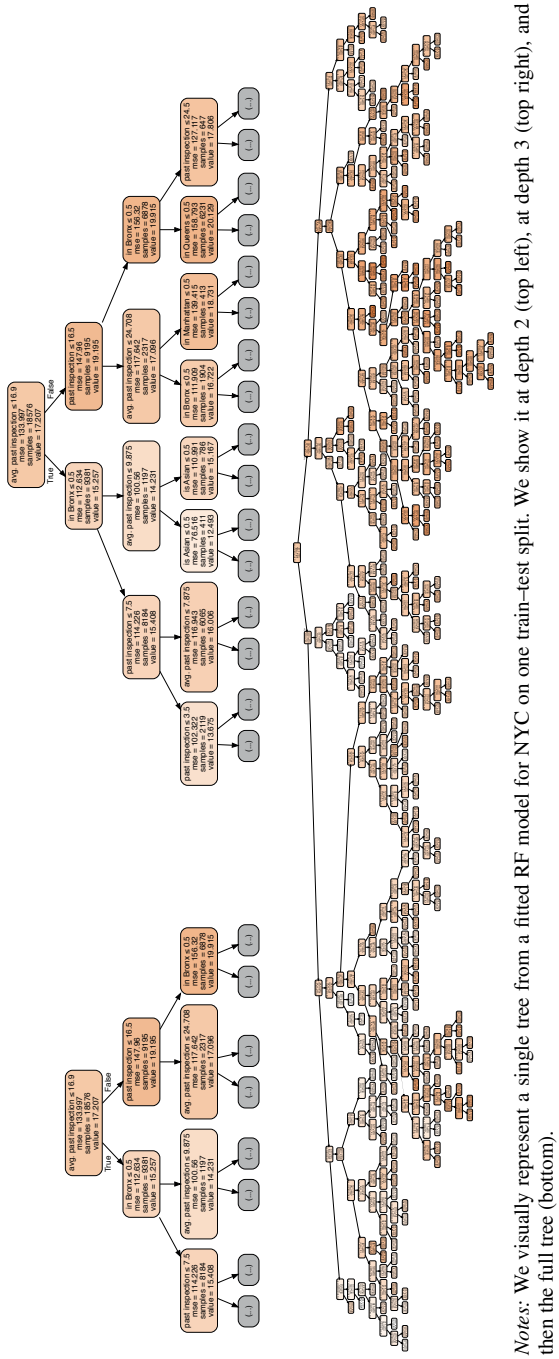
We provide a visual representation of a single decision tree from the fitted random forest models (on one train–test split) for both King County and New York. We represent the tree in increasing levels of depth (2, 3, full tree) for better readability.

Figure A7
A Single Tree from a Fitted Random Forest Model for King County



Notes: We visually represent a single tree from a fitted RF model for King County on one train–test split. We show it at depth 2 (top left), at depth 3 (top right), and then the full tree (bottom).

Figure A8
A Single Tree from a Fitted Random Forest Model for New York



Notes: We visually represent a single tree from a fitted RF model for NYC on one train–test split. We show it at depth 2 (top left), at depth 3 (top right), and then the full tree (bottom).