A Data-Driven Approach to Service Recommendations for Higher Recycling per Household in the State of Massachusetts

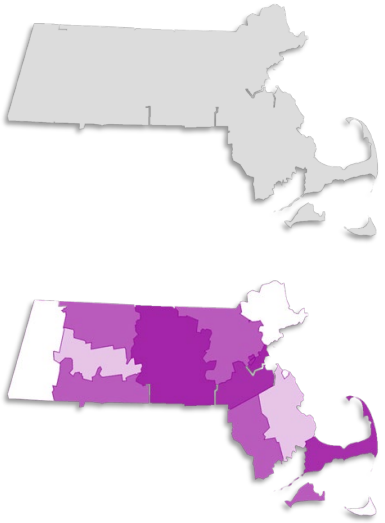
Certificate Program Study prepared by Diana Giulietti

Problem Statement and Scope

"What aspects of Municipal Waste Handling services do higher-recycling Massachusetts municipalities share?"

Scope

- Statistically may determine what features contribute the most to higher recycling rates per household
- Does **not** support that the services are the cause of higher recycling rates
  - Invisible influences include participant outlook and external political climate
- Analysis will be summarized on 5-years worth of data for the entire dataset and **for sub-sets of the data that represent rural, suburban and urban municipalities**
  - These sub-sets are derived from a clustering analysis



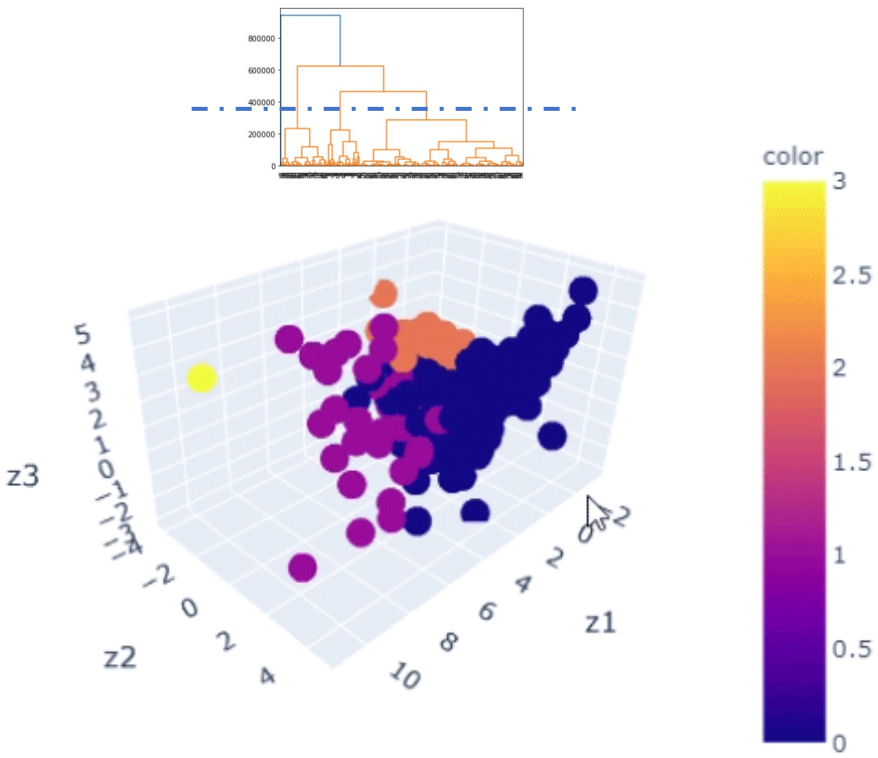
Data: MA Municipality Waste Handling Survey

Overview

- 350+ municipalities, 100+ question, Down-selected to 44 features
- Survey results for 2015-2019
  - Combined data for more robust dataset
- Includes information on
  - No. of Households served
  - Tonnage collected for trash and recycling
  - Trash and Recycling service types and volume of carts
  - Program funding mechanisms
  - Mandatory recycling, who it applies to, and who enforces it
- Many categorical features which required encoding

Clustering

Attempted Both KMeans Clustering and Hierarchical
Selected n = 4 based on analysis of dendrogram
Plotted results on PCA visualization of Census Attributes



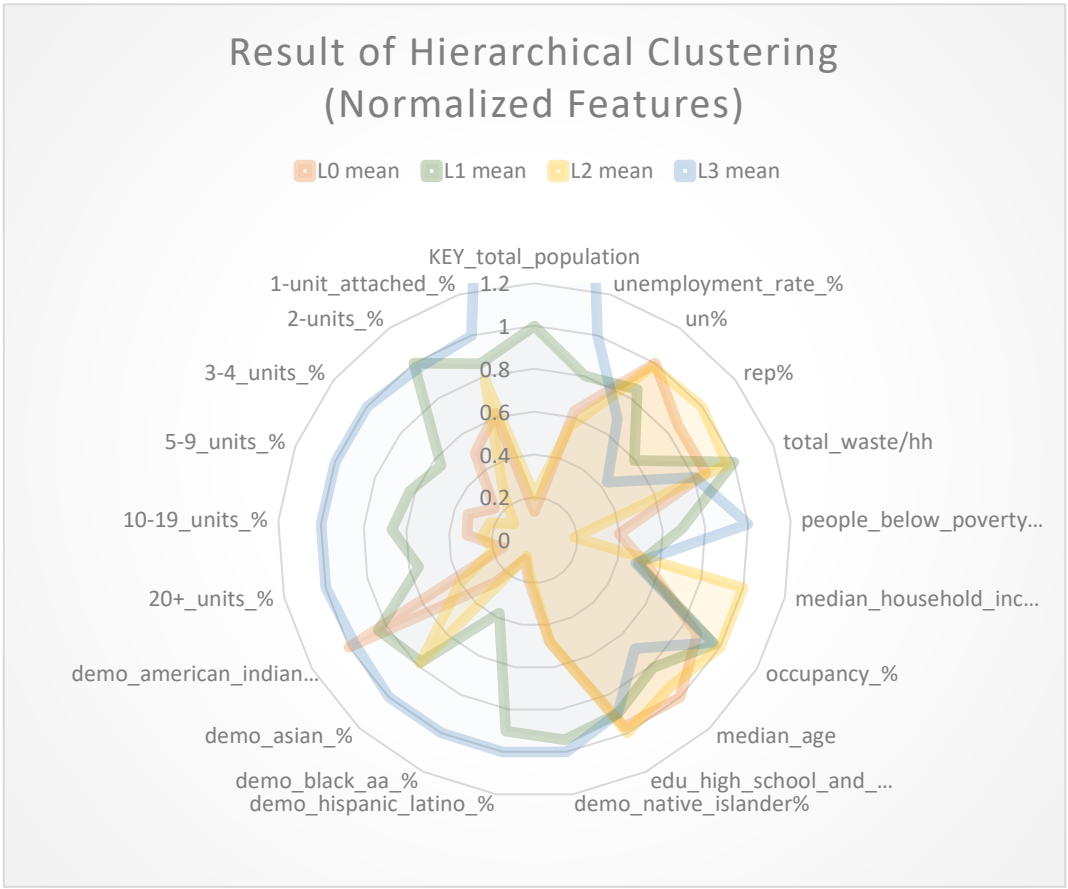
Labels roughly translate to:

"Rural" – Label 0, 179 municipalities

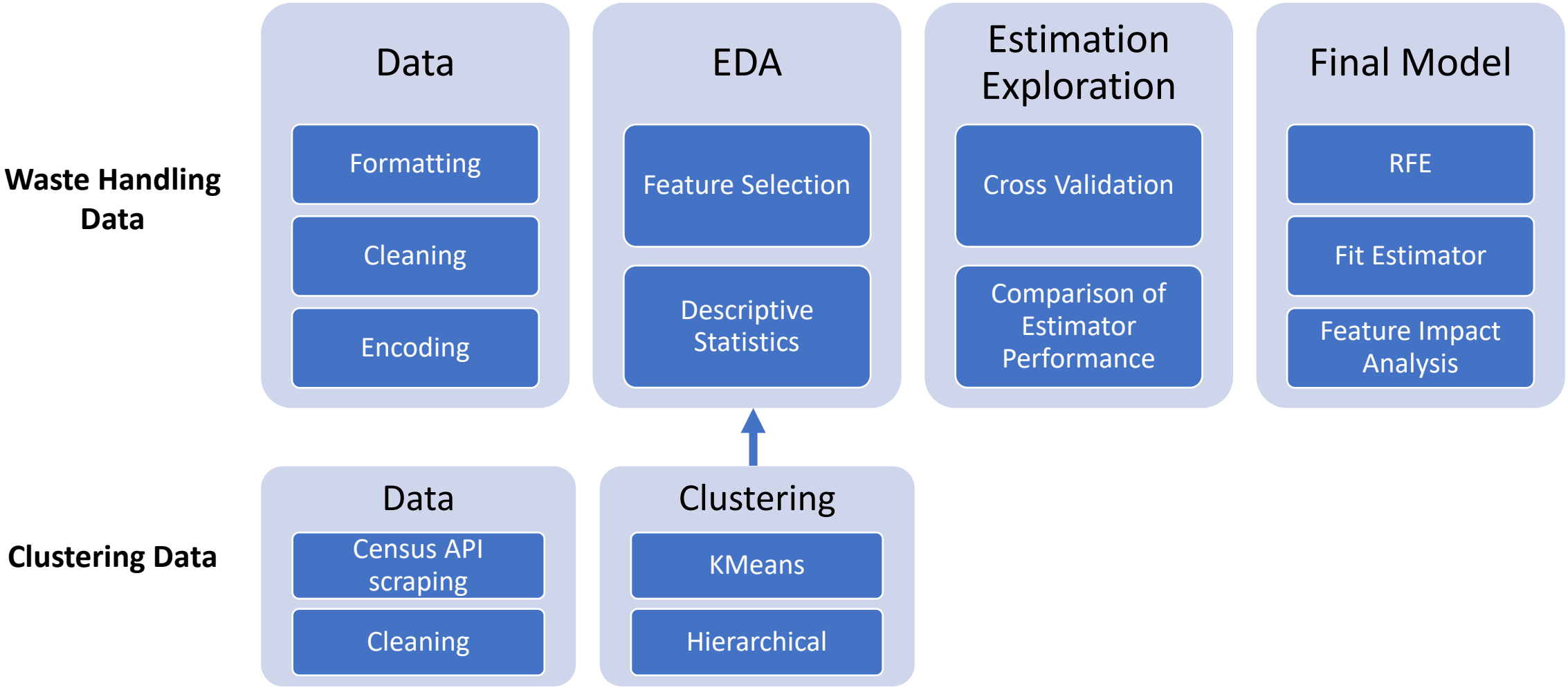
"Sub-urban" – Label 2, 52 municipalities

"Urban" – Label 1, 35 municipalities

Boston – Label 3, only 1 municipality

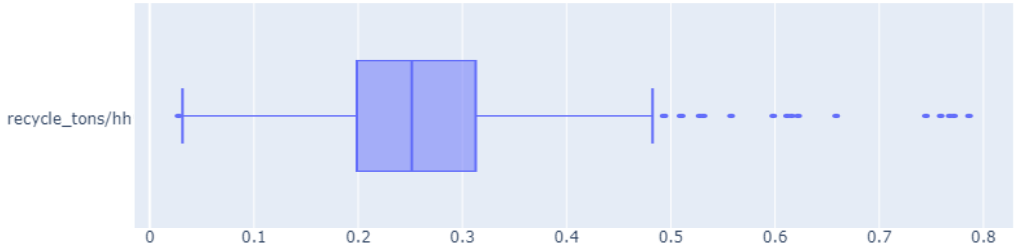


Data-Driven Process



Outliers: To include or not to include

Significant quantity of outliers but these data points may hold crucial information to maximizing recycling/household.



Performed Cross Validation to assess the robustness of various default estimators on the dataset with and without outliers

- Data appeared to not fit a linear function and non-parametric decision tree-based models produced the best results
- Non-parametric models scored similarly with and without outliers. Although standard deviation of the score was higher with outliers, the value of keeping the outliers in the model out-weigh the increase reproducibility.

All Data, Outliers Included			
	Linear Regression	**Random Forest**	**Bagging**
Fold1	0.141	0.444	0.388
Fold2	0.162	0.569	0.459
Fold3	0.186	0.600	0.550
Fold4	0.210	0.676	0.625
Fold5	0.086	0.565	0.450
CV_mean	0.157	0.571	0.494
CV_std	0.047	0.084	0.093

Trimmed Data, Outliers Excluded			
	Linear Regression	**Random Forest**	**Bagging**
Fold1	0.286	0.562	0.513
Fold2	0.286	0.576	0.563
Fold3	0.151	0.553	0.489
Fold4	0.151	0.523	0.453
Fold5	0.226	0.540	0.502
CV_mean	0.220	0.551	0.504
CV_std	0.067	0.020	0.040

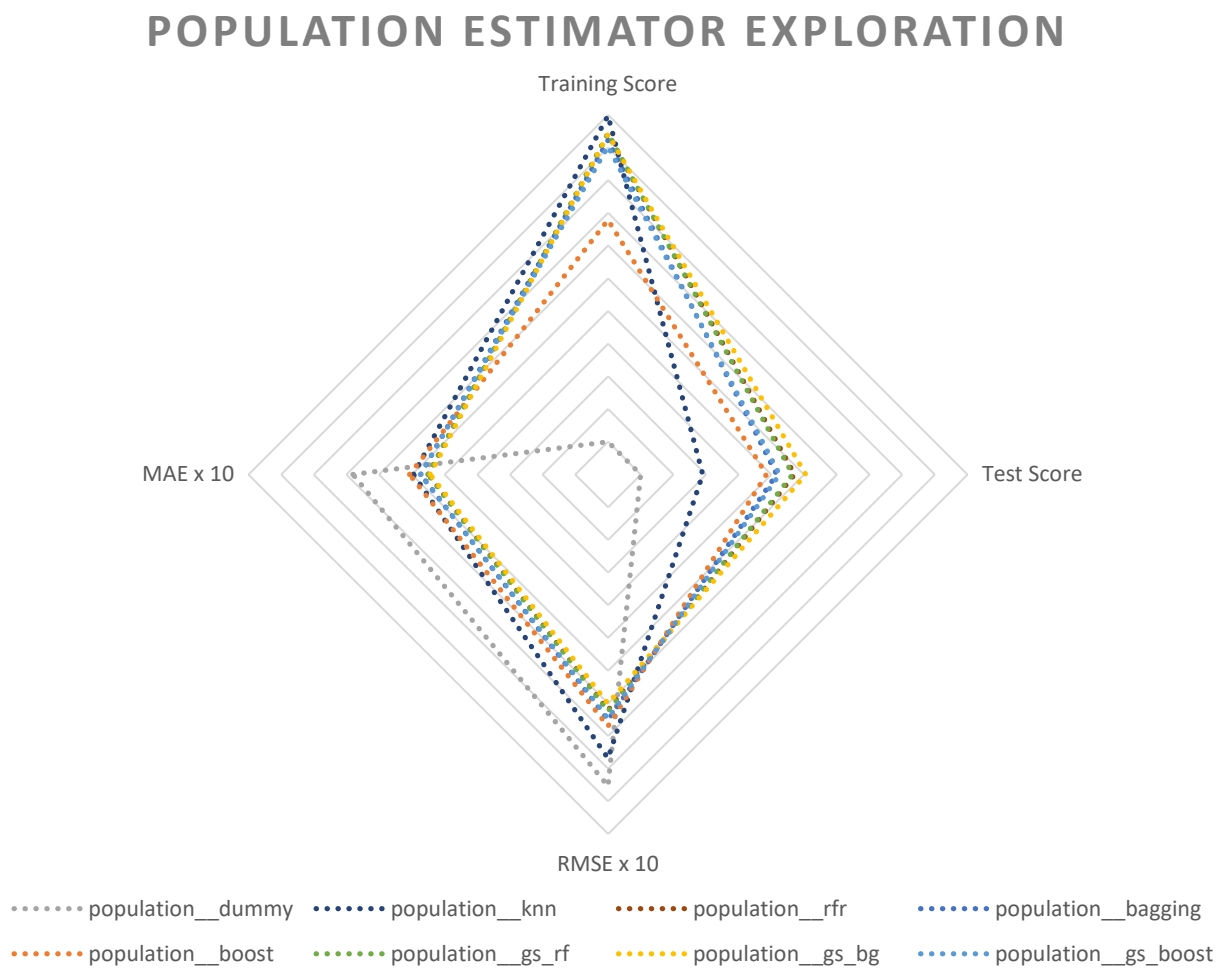
Estimator Selection: Cross Validation

Performed Cross Validation to assess the robustness of various default estimators on the dataset and **cluster subsets**

- Simple Random Forest and Bagging estimators
- L0 and L1 seems to have good and consistent fits when compared to the population
- L2 appeared to have one fold that significantly dropped the mean score and the increased the standard deviation (more on this when discussing model fit).

Estimator	Fold1	Fold2	Fold3	Fold4	Fold5	CV_mean	CV_std
population_RF	0.46	0.57	0.58	0.67	0.55	0.56	0.08
population_bagging	0.32	0.50	0.49	0.62	0.48	0.48	0.11
L0_RF	0.67	0.54	0.43	0.70	0.52	0.57	0.11
L0_bagging	0.59	0.46	0.33	0.63	0.52	0.51	0.12
L1_RF	0.41	0.65	0.68	0.60	0.58	0.58	0.11
L1_bagging	0.34	0.62	0.61	0.60	0.54	0.54	0.12
L2_RF	0.48	0.47	0.49	-0.30	0.26	0.28	0.34
L2_bagging	0.36	0.40	0.48	-0.41	0.29	0.22	0.36

Estimator Exploration : Population



Radar chart depicts the relative performance of various regression estimators ("gs" indicates these models are the result of hyperparameter tuning from GridSearchCV).

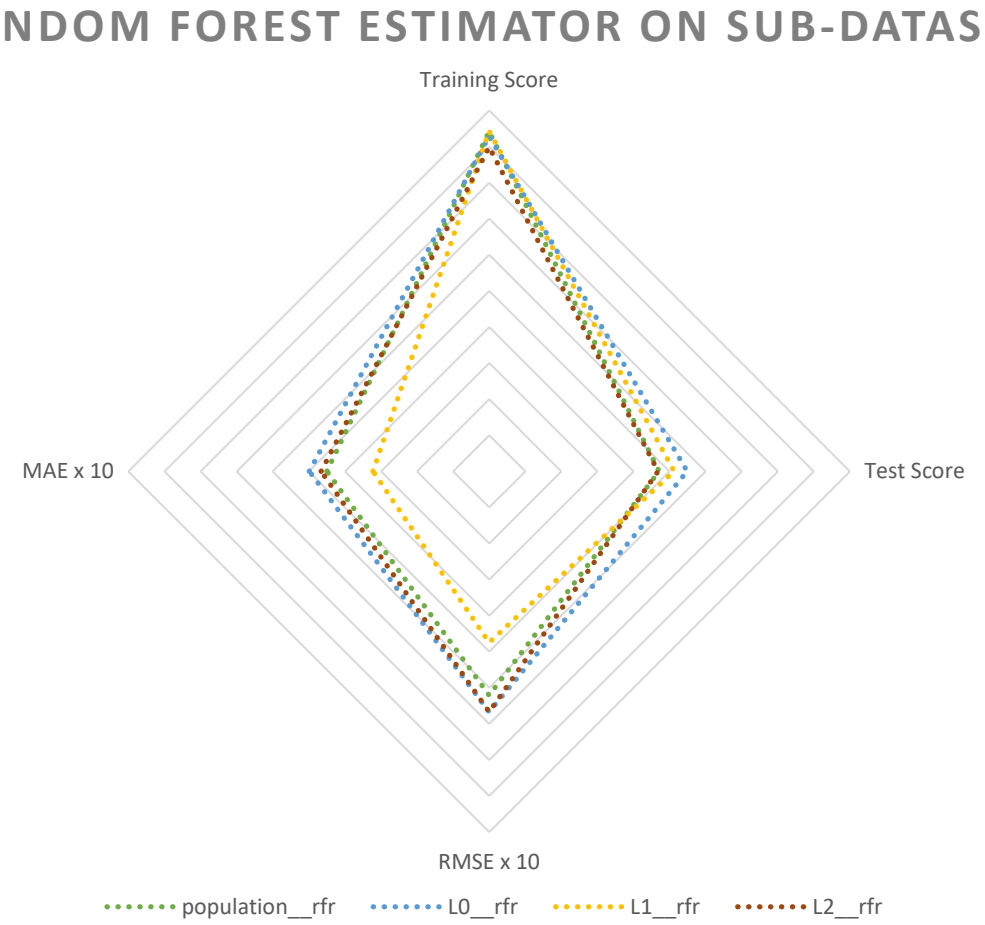
Dummy and kNN lazy estimator was used as a baseline models to assess performance against.

Random Forest and Bagging consistently performed well

- Test R^2 Scores around 0.45-0.50
- RMSE ≈ 0.60, MAE ≈ 0.45

Estimator Exploration : Cluster Subsets

- Clustered Data



Population	
score_train	0.944
score_test	0.463
rmse	0.063
mae	0.045
count	1258

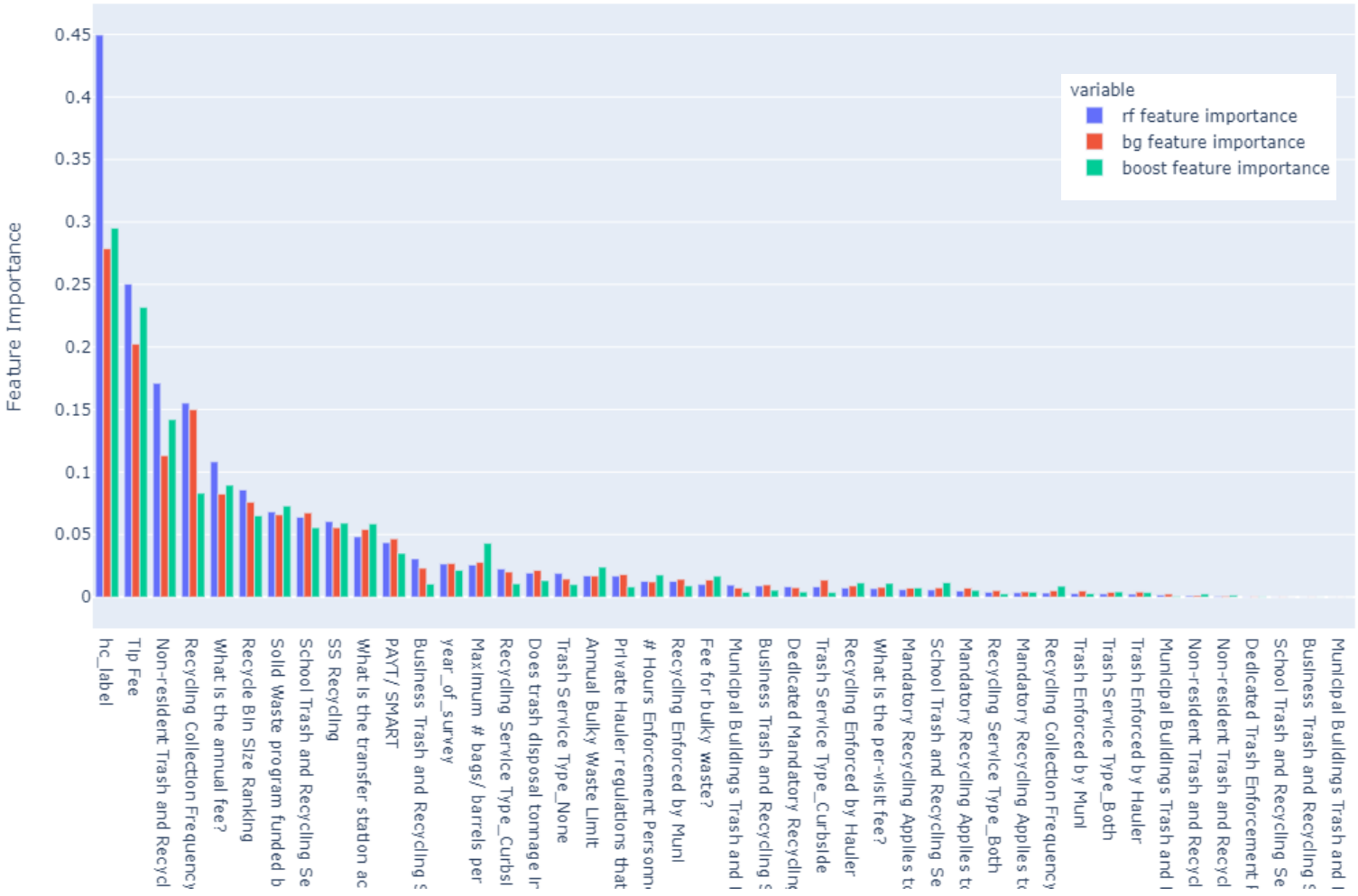
Rural (L0)	
score_train	0.927
score_test	0.556
rmse	0.066
mae	0.049
count	834

Urban (L1)	
score_train	0.941
score_test	0.530
rmse	0.047
mae	0.031
count	173

Suburban (L2b)	
score_train	0.907
score_test	0.469
rmse	0.067
mae	0.047
count	246

Important Features

Population



Feature importances (FI) calculated through permutation importance (leave out a feature and see how the model fit changes). FI represent how much the feature contributes to the model score *but does not reveal how data points will respond to this feature or provide actionable insights*.

Calculated for Random Forest, Bagging, and Gradient Boosting algorithms. Each model reflected similar rankings and values of feature importances.

Note: the cluster label feature (‘`hc_label`’) was found to be the most important feature, indicating that there may be distinct different in how the data performs in each one of the clusters. Also not that the year of the survey was included to ensure that this wasn’t unintentionally impacting trends. The importance of this feature is below 0.05, indicating low impact.

Model Fit and Feature Impact

Only showing top 5 impacting features (out of 22)

Population

Higher impact ⟷ Lesser impact

Municipality	True Tons Recycling /Household	Predicted Tons Recycling /Household	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact
29	0.616	0.505	Non-resident Trash and Recycling Service_Both	1	0.179	Recycling Collection Frequency_Weekly	1	0.033	Solid Waste program funded by property tax?	0	0.032	Maximum # bags/ barrels per week	0	-0.025	What is the transfer station access fee?	0	-0.019
173	0.509	0.496	Non-resident Trash and Recycling Service_Both	1	0.206	Maximum # bags/ barrels per week	2	0.031	Recycling Collection Frequency_Weekly	1	0.027	Tip Fee	37	-0.021	What is the transfer station access fee?	0	-0.02
228	0.531	0.180	hc_label	2	-0.061	Tip Fee	60.6	-0.014	Recycling Collection Frequency_Weekly	1	0.012	What is the annual fee?	0	-0.008	School Trash and Recycling Service_Both	0	-0.007
280	0.786	0.710	Non-resident Trash and Recycling Service_Both	1	0.207	What is the transfer station access fee?	30	0.046	Tip Fee	0	0.045	Maximum # bags/ barrels per week	1	0.031	Recycling Collection Frequency_Weekly	1	0.028
303	0.598	0.552	What is the annual fee?	242	0.201	year_of_survey	2016	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
422	0.622	0.561	Non-resident Trash and Recycling Service_Both	1	0.203	What is the transfer station access fee?	50	0.048	Tip Fee	55	-0.038	Maximum # bags/ barrels per week	2	0.037	Recycling Collection Frequency_Weekly	1	0.027
508	0.510	0.210	Tip Fee	71	-0.016	hc_label	0	-0.013	Recycle Bin Size Ranking	0	-0.009	Recycling Collection Frequency_Weekly	0	-0.009	School Trash and Recycling Service_Both	1	0.009
528	0.768	0.710	Non-resident Trash and Recycling Service_Both	1	0.207	What is the transfer station access fee?	30	0.046	Tip Fee	0	0.045	Maximum # bags/ barrels per week	1	0.031	Recycling Collection Frequency_Weekly	1	0.028
551	0.528	0.552	What is the annual fee?	254	0.201	year_of_survey	2017	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
679	0.558	0.561	Non-resident Trash and Recycling Service_Both	1	0.203	What is the transfer station access fee?	50	0.048	Tip Fee	57.55	-0.038	Maximum # bags/ barrels per week	2	0.037	Recycling Collection Frequency_Weekly	1	0.027
793	0.772	0.710	Non-resident Trash and Recycling Service_Both	1	0.207	What is the transfer station access fee?	30	0.046	Tip Fee	0	0.045	Maximum # bags/ barrels per week	1	0.031	Recycling Collection Frequency_Weekly	1	0.028
947	0.657	0.561	Non-resident Trash and Recycling Service_Both	1	0.203	What is the transfer station access fee?	50	0.048	Tip Fee	57.55	-0.038	Maximum # bags/ barrels per week	2	0.037	Recycling Collection Frequency_Weekly	1	0.027
1063	0.745	0.710	Non-resident Trash and Recycling Service_Both	1	0.207	What is the transfer station access fee?	30	0.046	Tip Fee	0	0.045	Maximum # bags/ barrels per week	1	0.031	Recycling Collection Frequency_Weekly	1	0.028
1087	0.759	0.552	What is the annual fee?	272	0.201	year_of_survey	2019	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
1217	0.612	0.549	Non-resident Trash and Recycling Service_Both	1	0.203	What is the transfer station access fee?	50	0.049	Tip Fee	93.75	-0.038	Maximum # bags/ barrels per week	2	0.037	Recycling Collection Frequency_Weekly	1	0.027

Gold text indicates that the prediction was more than 0.2 tons/hh off. The model may not be adequate in expressing these data points or these data points may have clerical errors, as was seen in serval outlier data points.

Backup

Model Fit and Feature Impact

Population

score_train	0.944
score_test	0.463
rmse	0.063
mae	0.045
count	1258

MinMaxScaler()

RFE(estimator=GradientBoostingRegressor(max_depth=5, random_state=8, subsample=0.66)

RandomForestRegressor(max_depth=20, n_estimators=1000, random_state=8)

Population's test set fit faired similar to the models fit in the estimator exploration step.

RMSE and MAE indicate acceptable levels of error (targeting the predicted variables to be within 0.05 tons/hh).

Feature Impact

To assess which features indicated high-recycling municipalities, the `treeinterpreter` library was used to 'white-box' the model. This library determines how a feature contributed quantitatively to the predicted dependent variable of an *individual data point.* A computationally intensive operation, only the top-performing outliers were selected; this is ultimately what would be important in determining the best recycling service configuration to maximize recycling per household.

Take-aways (*see follow slides for details*):

- Top performers have recycling and trash services for non-residential buildings
- Larger transfer station fees seemed to lead to better recycling (more funding for recycling services)
- The presence of a tip fee decreases recycling. This may first be counter intuitive but haulers are tipped on the precedence that they will check recycling bins to ensure the recycling in compliant before collecting. *The lower recycling rates reflects contaminated recyclables avoided.*

Model Fit and Feature Impact

Clustered Data

Population	
score_train	0.944
score_test	0.463
rmse	0.063
mae	0.045
count	1258

Rural (L0)

score_train	0.927
score_test	0.556
rmse	0.066
mae	0.049
count	834

MinMaxScaler()

RFE(estimator=GradientBoostingRegressor(max_depth=5, random_state=8, subsample=0.66)

RandomForestRegressor(max_depth=20, random_state=8)

Urban (L1)

score_train	0.941
score_test	0.530
rmse	0.047
mae	0.031
count	173

MinMaxScaler()

RFE(estimator=GradientBoostingRegressor(max_depth=5, min_samples_leaf=10, random_state=8)

RandomForestRegressor(max_depth=20, random_state=8)

Suburban (L2)

score_train	0.644
score_test	0.348
rmse	0.074
mae	0.055
count	246

MinMaxScaler()

RFE(estimator=GradientBoostingRegressor(max_depth=5, n_estimators=1000, random_state=8, subsample=0.33)

RandomForestRegressor(max_depth=5, max_samples=0.66, n_estimators=10, random_state=8)

Clusters all out-performed the population model with the exception of the L2 cluster. This L2 cluster underperformed even the models fitted in the estimator exploration stage. For this reason, a second model was fit, skipping the RFE step. The L2b model performed similarly to the population and to the models fit in the estimator exploration stage.

Suburban (L2b)

score_train	0.907
score_test	0.469
rmse	0.067
mae	0.047
count	246

MinMaxScaler()

RandomForestRegressor(max_depth=20, n_estimators=1000, random_state=8

Model Fit and Feature Impact

Clustered Data

Feature Impact Take-Aways (Details on Next Slides)

<u>Rural (L0):</u>

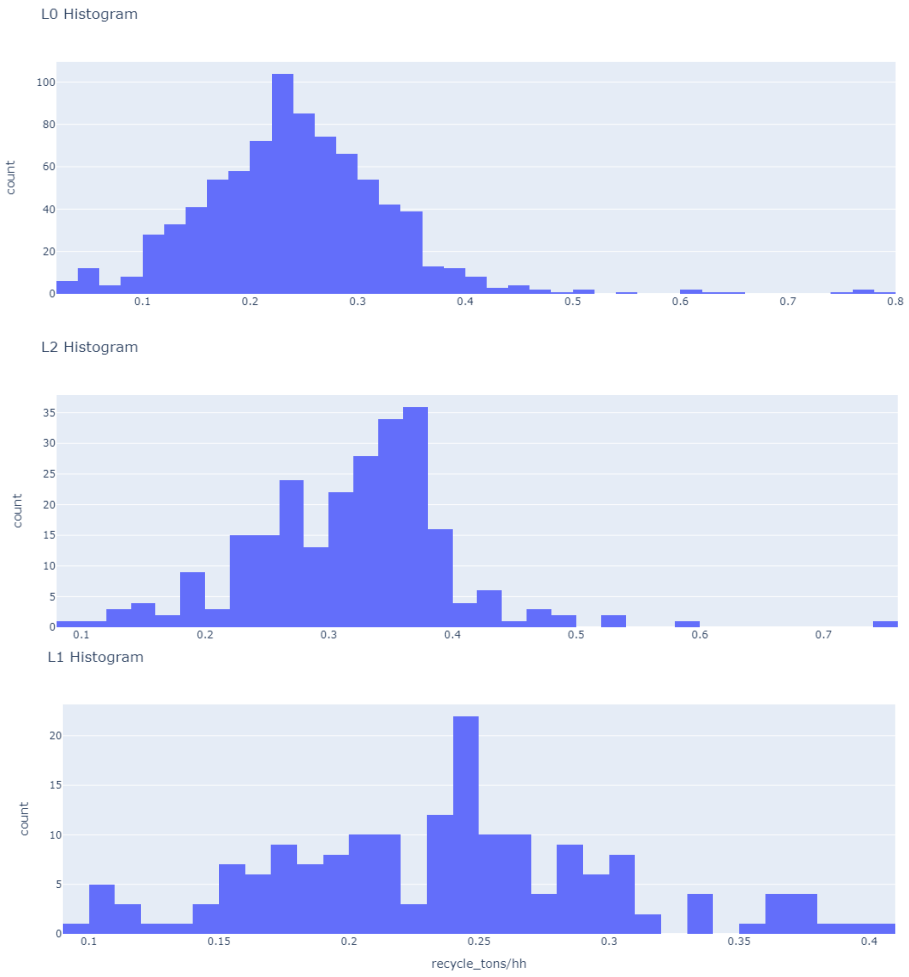
- Top performers have recycling and trash services for non-residential buildings
- Larger transfer station fees seemed to lead to better recycling (more funding for recycling services)
- Larger recycling bin (or no recycling limits) and weekly pickups yielded higher recycling per household
- Like the **Population Model**, the presence of a tip fee decreases recycling.

<u>Suburban (L2):</u>

- Focus was given to annual fees as a funding mechanism
- Tip fee resulted in increases in recycling as opposed to L0
- Favored weekly recycling pickup
- Seeing evidence of `year_of_survey` and `hc_label` being high-impact, suggesting that the other features were not significantly distinguishing. Could be the result of simply poor model fit or clerical errors in the data.
- **Model L2b** impact trends did not change much and still showed evidence of `year_of_survey` in the high impact variables. The predictions themselves were also further off than Model L2.
    - Further inspection of histograms of the dependent variable in each label shows that L2 is unique in that it is sharply skewed. This is likely causing abnormalities in the model fit as the test-train-split has limited points to pull from above 0.4 tons / household.

<u>Urban (L1)</u>

- Data subset only produced one outlier (histogram supports fairly gaussian behavior). More high performing points were sampled to look for trends.
- Favored recycling bin rank of 0.5 but did not favor a recycling bin rank of 1
- Positive response to tip fee unlike L1 and not funding the program with property tax responded better
- Supported recycling in business and schools lead to higher recycling and mandating recycling for private haulers also improved recycling per household.



Bins = 60

Model Fit and Feature Impact

Rural (L0)

Only showing top 5 impacting features (out of 22)

Higher impact ⟷ Lesser impact

Municipality	True Tons Recycling /Household	Predicted Tons Recycling /Household	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact
29	0.616	0.398	Non-resident Trash and Recycling Service_Both	1	0.127	Recycle Bin Size Ranking	1	0.031	Recycling Collection Frequency_Weekly	1	0.028	year_of_survey	2015	-0.026	Business Trash and Recycling Service_Both	0	0.019
136	0.463	0.222	Recycle Bin Size Ranking	1	0.028	Non-resident Trash and Recycling Service_Both	0	-0.008	Maximum # bags/ barrels per week	4	-0.011	SS Recycling	0	-0.016	Tip Fee	66	-0.022
173	0.509	0.447	Non-resident Trash and Recycling Service_Both	1	0.169	Recycle Bin Size Ranking	1	0.034	Business Trash and Recycling Service_Both	0	0.018	year_of_survey	2015	-0.017	Recycling Collection Frequency_Weekly	1	0.015
261	0.493	0.146	Non-resident Trash and Recycling Service_Both	0	-0.009	What is the annual fee?	40	-0.02	Recycle Bin Size Ranking	0	-0.022	Tip Fee	71	-0.048	Business Trash and Recycling Service_Both	1	0.004
280	0.786	0.609	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	30	0.061	Recycle Bin Size Ranking	1	0.033	Recycling Collection Frequency_Weekly	1	0.028	Business Trash and Recycling Service_Both	0	0.017
422	0.622	0.534	Non-resident Trash and Recycling Service_Both	1	0.178	What is the transfer station access fee?	50	0.063	Recycle Bin Size Ranking	1	0.034	Tip Fee	55	-0.028	Business Trash and Recycling Service_Both	0	0.018
508	0.510	0.146	Tip Fee	71	-0.048	Recycle Bin Size Ranking	0	-0.022	What is the annual fee?	40	-0.02	Non-resident Trash and Recycling Service_Both	0	-0.009	Business Trash and Recycling Service_Both	1	0.004
528	0.768	0.609	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	30	0.061	Recycle Bin Size Ranking	1	0.033	Recycling Collection Frequency_Weekly	1	0.028	Business Trash and Recycling Service_Both	0	0.017
679	0.558	0.553	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	50	0.065	Recycle Bin Size Ranking	1	0.034	Tip Fee	57.55	-0.027	Business Trash and Recycling Service_Both	0	0.018
664	0.452	0.379	Annual Bulky Waste Limit	156	0.104	Recycle Bin Size Ranking	1	0.031	Non-resident Trash and Recycling Service_Both	0	-0.008	SS Recycling	1	0.007	Tip Fee	86.1	0.004
793	0.772	0.609	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	30	0.061	Recycle Bin Size Ranking	1	0.033	Recycling Collection Frequency_Weekly	1	0.028	Business Trash and Recycling Service_Both	0	0.017
947	0.659	0.553	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	50	0.065	Recycle Bin Size Ranking	1	0.034	Tip Fee	57.55	-0.02	Business Trash and Recycling Service_Both	0	0.018
1063	0.745	0.609	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	30	0.061	Recycle Bin Size Ranking	1	0.033	Recycling Collection Frequency_Weekly	1	0.028	Business Trash and Recycling Service_Both	0	0.017
1217	0.612	0.553	Non-resident Trash and Recycling Service_Both	1	0.176	What is the transfer station access fee?	50	0.065	Recycle Bin Size Ranking	1	0.034	Tip Fee	93.75	-0.027	Business Trash and Recycling Service_Both	0	0.018

Gold text indicates that the prediction was more than 0.2 tons/hh off. The model may not be adequate in expressing these data points or these data points may have clerical errors, as was seen in serval outlier data points.

Model Fit and Feature Impact

Urban (L1)

Only showing top 5 impacting features (out of 22)

Higher impact ⟷ Lesser impact

Municipality	True Tons Recycling /Household	Predicted Tons Recycling /Household	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact
997	0.402	0.281	Recycle Bin Size Ranking	0.5	0.052	Tip Fee	59.09	0.018	Recycling Service Type_Both	0	-0.006	Private Hauler regulations that require recycling	0	-0.008	School Trash and Recycling Service_Both	1	0.006
1192	0.334	0.294	Recycle Bin Size Ranking	0.5	0.051	Tip Fee	68.97	0.01	Private Hauler regulations that require recycling	0	-0.01	Recycling Service Type_Both	1	0.009	year_of_survey	2019	-0.006
578	0.303	0.255	Maximum # bags/ barrels per week	1	0.015	Recycle Bin Size Ranking	1	-0.012	What is the annual fee?	0	-0.008	Tip Fee	0	0.007	School Trash and Recycling Service_Both	1	0.006
631	0.287	0.252	Recycle Bin Size Ranking	0	0.039	Tip Fee	0	-0.012	Business Trash and Recycling Service_Both	1	0.01	Private Hauler regulations that require recycling	0	-0.006	School Trash and Recycling Service_Both	1	0.005
74	0.319	0.292	Solid Waste program funded by property tax?	0	0.027	Business Trash and Recycling Service_Both	1	0.016	Private Hauler regulations that require recycling	1	0.015	Fee for bulky waste?	1	-0.009	# Hours Enforcement Personnel on Street	0	-0.008
845	0.32	0.224	year_of_survey	2018	0.012	Recycle Bin Size Ranking	1	-0.015	Trash Service Type_Both	1	0.008	Maximum # bags/ barrels per week	0	-0.008	What is the annual fee?	0	-0.008
535	0.382	0.335	Solid Waste program funded by property tax?	0	0.047	Business Trash and Recycling Service_Both	1	0.019	Private Hauler regulations that require recycling	1	0.014	# Hours Enforcement Personnel on Street	40	0.01	year_of_survey	2017	-0.007
1115	0.303	0.224	year_of_survey	2019	0.012	Recycle Bin Size Ranking	1	-0.015	Trash Service Type_Both	1	0.008	Maximum # bags/ barrels per week	0	-0.008	What is the annual fee?	0	-0.008
748	0.379	0.299	Recycle Bin Size Ranking	0.5	0.034	Tip Fee	108.09	0.016	Annual Bulky Waste Limit	52	0.014	Recycling Service Type_Both	1	0.009	Private Hauler regulations that require recycling	0	-0.009
511	0.337	0.309	Recycle Bin Size Ranking	0.5	0.027	Private Hauler regulations that require recycling	1	0.021	year_of_survey	2017	-0.01	Annual Bulky Waste Limit	52	0.009	Solid Waste program funded by property tax?	0	0.008
1017	0.289	0.279	Recycle Bin Size Ranking	0.5	0.034	Tip Fee	108.09	0.016	Annual Bulky Waste Limit	52	0.014	Private Hauler regulations that require recycling	0	-0.008	year_of_survey	2018	-0.007

Top data point (977) is the only outlier in L1. The remaining 10 points were randomly sampled from the top quartile of the subset.

Model Fit and Feature Impact

Suburban (L2)

Higher impact ⟷ Lesser impact

Only showing top 5 impacting features (out of 22)

Municipality	True Tons Recycling /Household	Predicted Tons Recycling /Household	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact	Feature	Value	Impact
228	0.531	0.180	hc_label	2	-0.061	Tip Fee	60.6	-0.014	Recycling Collection Frequ	1	0.012	What is the annual fee?	0	-0.008	School Trash and Recycling Service_Both	0	-0.007
303	0.598	0.552	What is the annual fee?	242	0.201	year_of_survey	2016	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
551	0.528	0.552	What is the annual fee?	254	0.201	year_of_survey	2017	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
1087	0.759	0.552	What is the annual fee?	272	0.201	year_of_survey	2019	0.028	hc_label	2	0.016	Recycling Collection Frequency_Weekly	1	0.015	Tip Fee	79	0.009
228	0.531	0.277	Recycling Collection Frequency_Weekly	1	0.01	Tip Fee	60.6	-0.031	Business Trash and Recycling Service_Both	1	-0.01	What is the annual fee?	0	-0.006	Does trash disposal tonnage include bulky waste?	1	-0.006
303	0.598	0.454	What is the annual fee?	242	0.092	year_of_survey	2016	-0.019	Tip Fee	79	0.012	Recycling Collection Frequency_Weekly	1	0.011	SS Recycling	0	0.011
551	0.528	0.454	What is the annual fee?	254	0.092	year_of_survey	2017	-0.019	Tip Fee	79	0.012	Recycling Collection Frequency_Weekly	1	0.011	SS Recycling	0	0.011
1087	0.759	0.454	What is the annual fee?	272	0.092	year_of_survey	2019	-0.019	Tip Fee	79	0.012	Recycling Collection Frequency_Weekly	1	0.011	SS Recycling	0	0.011

Gold text indicates that the prediction was more than 0.2 tons/hh off. The model may not be adequate in expressing these data points or these data points may have clerical errors, as was seen in serval outlier data points.

Top four rows show results from L2 Model. Bottom four show results from L2b Model with the same points.