Stinky NYC

– A Machine Learning Approach to Explain NYC Urination or Defecation Complaints

Fall 2022 Machine Learning for Cities Final Project

December 2022

Jingjing Ge jg5788@nyu.edu | Yichen Guo yg1835@nyu.edu

Jiashun Lian jl14414@nyu.edu | Chaofan Zheng cz2758@nyu.edu

Github Link: https://github.com/jingjingge/MLC_FInal_Project_Group2

1. Abstract

The issue of public urination or defecation in New York City arose a decade ago. In the summer of 2022, the New York City odor complaint reached an all time high with people complaining about the city smell like used diapers[1]. While public urination and defecation not only triggers disgusting scent, it also brings hygienic problems to the public. Based on our searches, the city hasn't brought up a concrete and effective solution to ease the burden. This project utilized the 311 Complaints data about public urination and defecation to find patterns of areas with high complaints using machine learning methods. We classified the city areas into high and low urination occurrences. Our best model had an accuracy of 90%, and could be used to provide insights for better public urination and health management.

Key Words: Urination density, Machine Learning, Regression, Tree-like Model

2. Introduction

2.1 How New York City react to public urination or defecation

Before 2016, public urination or defecation was a misdemeanor under Public Health charge, and would leave a criminal record[2]. The Criminal Justice Reform Act of 2016 then 'downgraded' the punishment of public urination or defecation into civil penalties due to the so-called 'quality of life'[3]. In recent years, the city of New York has implemented measures to try to reduce public urination, such as installing additional public restrooms and increasing fines for individuals caught urinating in public. Despite these efforts, the issue remains a persistent problem in the city as complaints keep rising.

2.2 Related works

While we didn't encounter any machine learning method based approach towards the issue of public urination or defecation, an exploratory analysis from Kaggle indicates that these complaints are correlated to warm weather conditions and are more likely to occur during working hours on weekdays[4].

The Kaggle analysis was conducted way back to 2015 and there didn't appear to be any data driven approach targeting public urination and defecation since then.

Although we lack precise work directly related to public urination or defecation, there is plenty of machine learning based research on the NYC 311 complaints and NYC Open Data. For example, this paper by Ransome et.al deployed a Bayesian hierarchical Poisson regression to focus on the relationship between alcohol-related complaints and alcohol outlet density, area-level drinking and sociodemographic factors[5]. Another project focuses on classification of whether a crime is successful or unsuccessful[6]. This project used the NYPD Complaint Data and trained logistic regression, multi-layer perceptron, decision tree and random forest on it, with an optimal accuracy of 98%. Based on previous works, we found it reasonable and feasible to apply machine learning methods to target NYC's urination issue.

3. Data

3.1 Data Collection

The unit of analysis in this project is zip codes since earlier research in NYC considered zip codes reasonable proxies for neighborhoods and due to the availability of data (not all datasets were available at a smaller geographical level.)

This research mainly used open data from government websites and databases. We used the public urination complaints data as the main training, validation and testing sets for the models along with other features. The data was extracted from 311 service requests from 2010 to present provided by 311 and DoITT and is publicly available on NYC Open Data. The geographical level we used to aggregate and classify the complaints in New York CIty was based on zip code boundaries, for which the shapefiles were collected from the Department of City Planning official website, and all the other features were also collected in zip code level for later data merging. The socioeconomic and demographic data (age, income, race, sex, education level) were collected from ACS 5-year census estimate by calling API on United Census Bureau website. Neighborhood features such as NYPD complaints, number of subway entrances,

green space area and street tree census were also collected from NYC Open Data provided by various government agencies. Other features including the public restroom and commercial shops were exported from the OpenStreetMap using QGIS.

3.2 Data Cleaning and Processing

The public urination complaint data was extracted from the 311 service requests by filtering the "Complaint Type" column to "urinating in public". As there isn't a huge number of public urination complaints in recent years, we decided to keep all the data from 2010 to present, which are in total 7720 rows. Since each row of the complaint has the attribute of zip code, they were simply further aggregated and normalized by population for each zip code area. These areas are later classified as high or low occurrence of urination by number 0 and 1 based on the complaint density calculated before. In order to merge all the features data into one dataset based on the zip code, we used spatial join on the geopandas to count the number of features (crime, subway entrances, trees, shops, restroom, etc) in each zip code area and some were further normalized by the area or population based on the characteristics of the feature to be unbiased. We only used recent two years of NYPD complaints data instead of starting from 2010 because the numbers are too large for processing, so we assume that the crime situation hasn't changed much in the last 10 years. For the demographics datasets, we dropped some zip codes as they have missing census data because the area is too small or impossible for people to settle. Last but not least, we merged all datasets together in one dataset with 167 rows representing the zip code area and 57 columns representing the features and dropped any missing values for one more time. The final cleaned data features are attached in the appendix B.

4. Methods

4.1 Regression Model

4.1.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best fitting straight line that describes the relationship between the dependent and independent variables.

4.1.2 Lasso for Feature Selection

Lasso regression, or Least Absolute Shrinkage and Selection Operator, is a variant of linear regression. One reason to use a Lasso model for variable selection is that it can be used to reduce the number of variables and eliminate multicollinearity, which is especially useful for high-dimensional data and to increase model interpretability[7]. We first normalized the data to fit into a Lasso regression. Then the features selected from our Lasso regression were then fitted into the linear regression as mentioned above.

4.2 Classification Model

We use the K-means algorithm to divide New York into two clusters based on the density of urination in different regions. Our original K-means model resulted in three clusters, however, there were only 6 samples in the third cluster, which would not benefit the model training process. We also observed the distribution of data, determining 2 clusters would be enough in separating the samples properly with class 0 representing low urination complaint area and class 1 representing high urination complaint area. We fine-tuned the classification models using RandomSearch since GridSearchCV was time consuming.

4.2.1 SVM Models

Support vector machine (SVM) is a supervised machine learning model that is commonly used for classification and regression tasks. It seeks to find the decision boundary that optimally separates the different classes in the training data. Figure below shows our choices of fine-tuning parameters for SVM.

Parameters	Ranges	Optimal values
kernel	[Linear, rbf]	rbf
C	[0.001, 0.01, 0.1, 1, 10, 100]	10
gamma	[0.001, 0.01, 0.1, 1, 10, 100]	0.001

Figure 1: Fine tuning setups for SVM

4.2.2 Tree-based Models

We used tree based models since they are not sensitive to feature collinearity .They do not rely on linear relationships between the features and the target variable. Instead, they partition the feature space into regions based on the values of the features, and make predictions based on which region a new sample falls into.

Decision Tree (DT)

Decision tree model is a supervised learning algorithm that uses a tree-like structure to make predictions. It can be understood and interpreted easily, and it can handle both numerical and categorical data. Figure below shows our choices of fine-tuning parameters for the decision tree.

Parameters	Ranges	Optimal values
splitter	['best', 'random']	best
max_features	[None, 'sqrt', 'log2', 0.2, 0.4, 0.6, 0.8]	sqrt
max_depth	[None, 2, 4, 6, 8, 10]	4
criterion	['gini', 'entropy']	gini

Figure 2: Fine tuning setups for DT

Random Forest (RF)

Random forest models use multiple decision trees to predict data. It has two main advantages: resistant to overfitting and can be trained in parallel, which improves training efficiency. Figure below shows our choices of fine-tuning parameters for our Random forest model.

Parameters	Ranges	Optimal values
n_estimators	[100, 120, 150]	100
min_samples_split	[2, 3,···, 10]	9
min_samples_leaf	[1,3,5]	1
max_features	[1, 2,···, 10]	7
max_depth	[None,1,3,5,7,9]	5
criterion	['entropy', 'gini']	entropy

Figure 3: Fine tuning setups for RF

XGBoost

XGBoost is an implementation of gradient boosting, which is an ensemble learning method that combines multiple weak learners to form a strong learner. One of the main advantages of XGBoost is that it uses a regularization term to control overfitting. Another advantage of XGBoost is that it uses a technique called sparsity-aware splitting, which helps to reduce the number of splits in the decision trees and make the model more interpretable. Figure below shows our choices of fine-tuning parameters for XGBoost.

Parameters	Ranges	Optimal values
n_estimators	[100, 250, 500, 750]	750
learning_rate	[0.01, 0.1, 0.2, 0.3, 0.4]	0.3
max_depth	[3, 6, 10, 15]	6
subsample	[0.5, 0.6,···, 1]	0.6
tree_method	[auto, exact, approx, hist]	approx
colsample_bytree	[0.5, 0.6,···, 1]	0.9
colsample_bylevel	[0.5, 0.6,···, 1]	0.7
objective	['binary:logistic','binary:logitraw']	binary:logitraw

Figure 4: Fine tuning setups for XGBoost

4.3 SHAP (SHapley Additive exPlanations) Explainer

To increase our tree models transparency and interpretability, we deployed the SHAP explainer. According to SHAP documentation, SHAP uses a game theoretic approach that can explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions[8]. We looked into both SHAP scatter plot and SHAP summary plot of mean absolute SHAP value.

In our scenario, a negative SHAP value means the feature is prone to the lower urination complaint class. The more negative the value is, the larger weight it would have on the prediction of lower complaint class. Vice versa, a positive SHAP value means the feature is prone to the higher urination complaint class. The more positive the value is, the larger weight it would have on the prediction of higher complaint class.

5. Results

5.1 Regression Model

5.1.1 Lasso Model for feature selection

According to the processing results, the optimal alpha value was selected to be 0.05, in order to prevent overfitting due to small alpha. To reduce multicollinearity in the latter linear regression process, the optimal five factors were selected. The five factors we separated out first were 'Associate Degree','subway count', 'Percent Age Under 5', ' high school Graduate', 'green space ratio'.

5.1.2 Linear Regression

OLS Regression Results

Dep. Variable:	urination_density	R-squared:	0.480
Model:	OLS	Adj. R-squared:	0.460
Method:	Least Squares	F-statistic:	23.99
Date:	Sat, 10 Dec 2022	Prob (F-statistic):	4.24e-11
Time:	06:46:57	Log-Likelihood:	978.68
No. Observations:	82	AIC:	-1949.
Df Residuals:	78	BIC:	-1940.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.037e-06	7.53e-07	5.358	0.000	2.54e-06	5.54e-06
subway_count	4.768e-08	2.19e-08	2.181	0.032	4.15e-09	9.12e-08
Associate_Degree	-4.783e-05	8.78e-06	-5.447	0.000	-6.53e-05	-3.03e-05
green_space_ratio	2.604e-06	1.01e-06	2.569	0.012	5.86e-07	4.62e-06

Omnibus:	34.214	Durbin-Watson:	2.017
Prob(Omnibus):	0.000	Jarque-Bera (JB):	106.988
Skew:	1.293	Prob(JB):	5.86e-24
Kurtosis:	7.962	Cond. No.	765.

Figure 5: OLS regression results

The variables were tuned according to iterations to achieve a higher fit at a lower covariance. The above regression results were obtained. Controlling Cond. No. for fewer numbers to achieve the effect of avoiding a high degree of covariance. The in-sample model fits with an R squared of 0.480. There was a slightly negative correlation between ‘Associate Degree’ and the dependent variable, while a slightly positive correlation with other variables. Calling the test set and comparing it with the predicted values, an MSE of 1.7216e-06 was obtained, with a predicted accuracy of 0.446140.

5.2 Classification Model

Figure below shows a summary of our classification model performance. Most of our models achieved scores over 0.8 in all metrics, including Precision, Recall, F1 and Accuracy. We used the most basic decision tree model as a benchmark, and the remaining three models all had over 10 percentage improvements. Among all the models, XGBoost had the best predictive performance, achieving 0.9 on Accuracy and Precision, 0.89 on Recall and F-1score (figure 6).

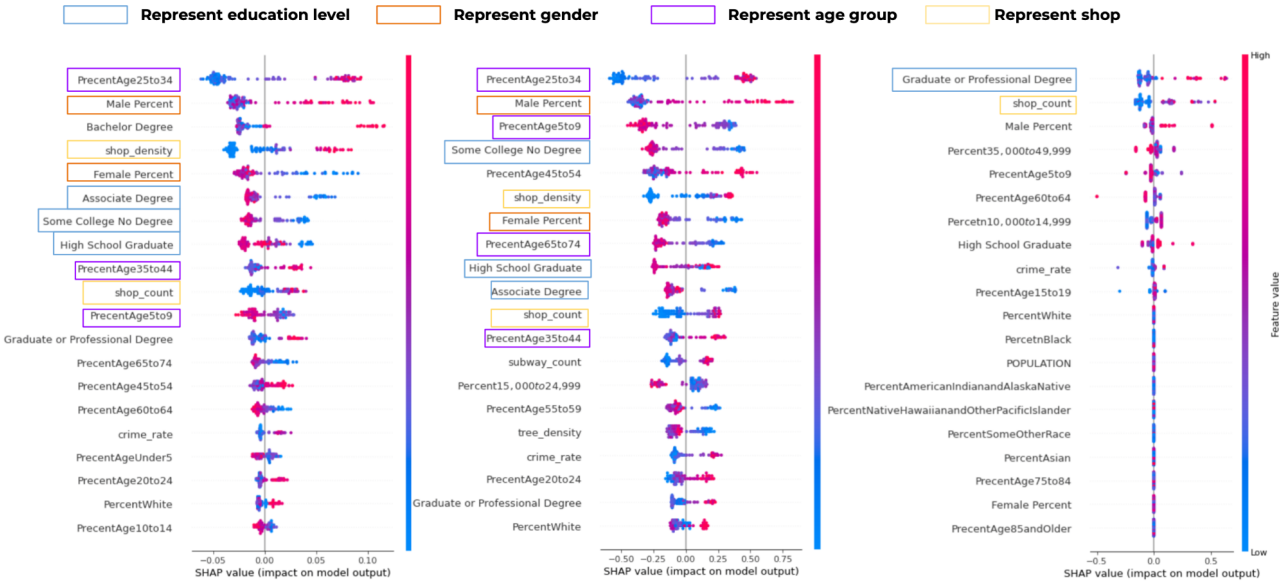
Model		Precision	Recall	F1-score	Accuracy	Fine Tuning time(s)	Fit time(s)	Support
Decision Tree	class 0	0.8	0.88	0.83	0.79	1.39	0.101	40
	class1	0.77	0.65	0.71				26
	weighted avg	0.79	0.79	0.78				66
Random Forest	class 0	0.83	0.97	0.9	0.86	104	0.324	40
	class1	0.95	0.69	0.8				26
	weighted avg	0.88	0.86	0.86				66
SVM	class 0	0.91	0.91	0.91	0.88	5.03	0.09	46
	class1	0.8	0.8	0.8				20
	weighted avg	0.88	0.88	0.88				66
XGBoost	class 0	0.87	0.97	0.92	0.9	183	1.45	40
	class1	0.95	0.77	0.85				26
	weighted avg	0.9	0.89	0.89				66

Figure 6: Classification models summary

5.3 SHAP (SHapley Additive exPlanations) Explainer for Feature Importance

Figure 7 shows the feature importance from SHAP in the scatter summary plot. The feature importance distribution had higher similarities between the RF and XGBoost SHAP result. There were more features with negative SHAP value among all zip codes, which are represented by more concentrated scattered points in the plot. More specifically, the lower complaint tendency was associated with lower percent of male and the population between age 25 and 34, less number of shops, higher percent of population with at least a high school degree, higher percentage of population between age 5 and 9 and age 65 and 74.

The summary plot of mean absolute SHAP value reveals the features that played important roles in our model prediction regardless of clusters. The relatively significant features are: percentage of population between 25 and 34, percentage of male, shop density, percentage of population from age 5 to 9, percentage of population with at least high school education (associate degree, some college no degree, high school degree).



Random forest SHAP summary XGBoost SHAP summary Decision tree SHAP summary

Figure 7: SHAP summary scatter plot

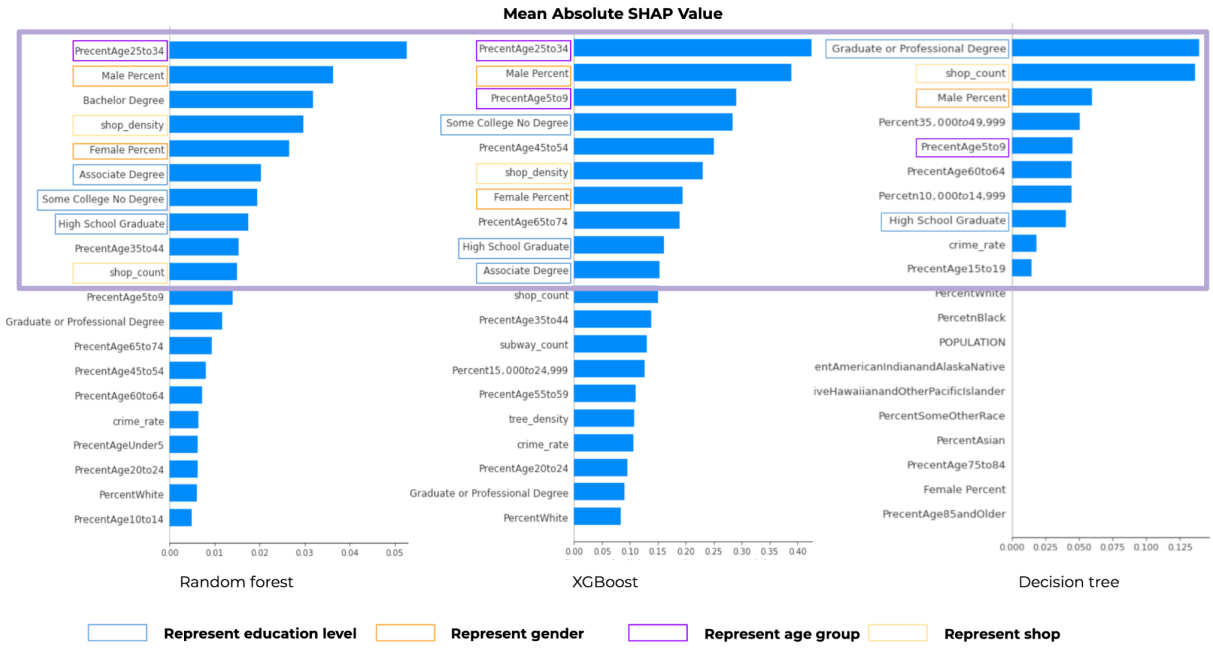


Figure 8: Mean absolute SHAP value summary

6. Conclusion

6.1 Model Performance

The reason that Random Forest performs better than Decision Tree is that it can absorb more information from multiple decision trees and effectively reduce the risk of overfitting. In addition, the Random Forest model will be trained multiple times by random sub-sampling of features during the training process, which helps to improve the generalization ability of the model.

SVM can usually perform better on small-sample high-latitude imbalanced datasets because it can effectively deal with nonlinear data, and can map high-dimensional data to lower-dimensional spaces through kernel functions, thereby avoiding disasters caused by the curse of dimensionality. resulting in performance degradation. In addition, SVM can also solve the problem of sample imbalance by assigning different weights to each sample point, thereby improving the generalization ability of the model.

The XGBoost model can use various regularization items to control the complexity of the model, and achieve better model performance by adjusting the depth of the tree and the splitting conditions of the leaf nodes. These properties allow XGBoost models to converge faster and in most cases achieve higher accuracy thanRandom Forest and SVM..

6. 2 Feature Findings

In conclusion, we found that the classification model is more suited for our project and produces better results than linear regression. However, the lasso regression presents some useful insights on the feature importance. It shows that green space ratio and subway count are positively correlated with the number of urination complaints, and the percentage of people with associate degree, high school graduate, and age under 5 are negatively correlated with number of complaints. This result overlaps with our classification model as the SHAP in the previous section explained that age, gender, education level and shop are the most correlated features. The income level has less correlation since it is only presented in

the DT model. We can infer from these features that younger people from 25-34 may tend to file more complaints, while the elders may not since they are not familiar with the 311 system. Parents with kids from age 5-9 tend to choose better living environments that have less urination complaints. Areas that have less shops have lower complaints because there are less people in the area. Lastly, well educated areas also tend to have less urination complaints since they are more aware of public health issues.

Although we have some interesting findings, we cannot say that these features are always correlated with the occurrence of the urination. Since our data is only based on people who chose to report the urination incident, there might be more urination happening in vacant land or ignored by people. Therefore, we could only say that these features are related to a higher number of complaints. In order to have more data, the city government could make the 311 more accessible to the public by encouraging people to use the app such as giving some incentives.

6.3 Future Considerations

Since we made predictions based on zip code in NYC, which was a relatively small sample size. In the future, we may consider looking into smaller scale data like census tract. We may also gather yearly data rather than aggregated 10-year public urination data to gain more training samples. It would also be interesting to perform a time series analysis on the seasonal trends of the complaints. By adding weather and date-time data, we would have a better understanding of the trends of the complaints, so proper measures can be taken based on the seasonal trends observed. We can also try to figure out what is the major source of odor related complaints by adding other complaints such as odor from trash, food waste, animal waste and chemicals. We also observed that the amount of complaints rose since the lock down of COVID, it would also be insightful to take a specific look at what happened during the pandemic time. Also, since we have a small sample size, we can do oversampling to expand the sample size for better model performance.

References

[1] Aidalalaw. “Is New York City about to Ease up on Petty Crimes?” *Aidala, Bertuna & Kamins – New York City Trial Attorneys*, 9 Aug. 2021, https://aidalalaw.com/is-new-york-city-about-to-ease-up-on-petty-crimes/.

[2] “Public Urination Health Code Statute Text and Explanation.” *Pink Summons Information*, https://www.pinksummons.com/public-urination-health-code-summons-nyc.

[3] Aidalalaw. “Is New York City about to Ease up on Petty Crimes?” *Aidala, Bertuna & Kamins – New York City Trial Attorneys*, 9 Aug. 2021, https://aidalalaw.com/is-new-york-city-about-to-ease-up-on-petty-crimes/.

[4] Jasonduncanwilson. “Urination in NYC and Other Fun Exploration.” *Kaggle*, Kaggle, 18 May 2018, https://www.kaggle.com/code/jasonduncanwilson/urination-in-nyc-and-other-fun-exploration/notebook.

[5] Ransome, Y., Luan, H., Shi, X. et al. Alcohol Outlet Density and Area-Level Heavy Drinking Are Independent Risk Factors for Higher Alcohol-Related Complaints. J Urban Health 96, 889–901 (2019). https://doi.org/10.1007/s11524-018-00327-z

[6] Htappa. “NYC_CrimeData/NYPD_CRIMEDATA.Ipynb at Master · Htappa/NYC_CRIMEDATA.” *GitHub*, https://github.com/htappa/NYC_CrimeData/blob/master/NYPD_CrimeData.ipynb.

[7] Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1-25.

[8] “Welcome to the Shap Documentation.” *Welcome to the SHAP Documentation - SHAP Latest Documentation*, https://shap.readthedocs.io/en/latest/.

Team contribution

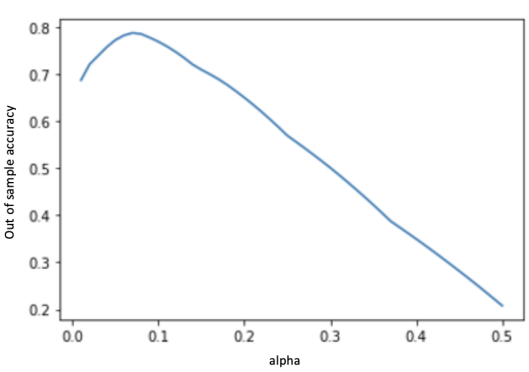
Jingjing Ge: Topic research; Data collection, cleaning, aggregation; Model discussion; Presentation; Paper write ups

Yichen Guo: Topic research; Data collection, cleaning, aggregation; Model discussion; Presentation; Paper write ups

Jiashun Lian: Topic research; Data suggestion; Model preparation, realization, tuning; Presentation; Paper write ups

Chaofan Zheng: Topic research; Data suggestion; Model preparation, realization, tuning; Presentation; Paper write ups

Appendix A



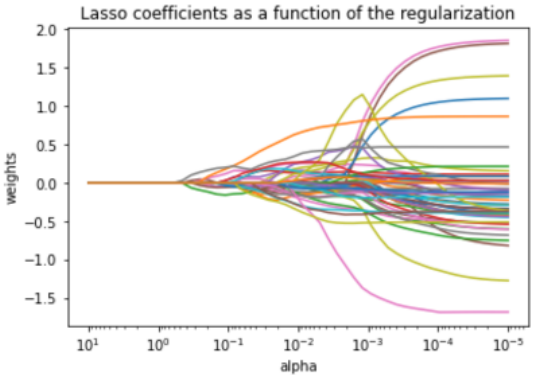
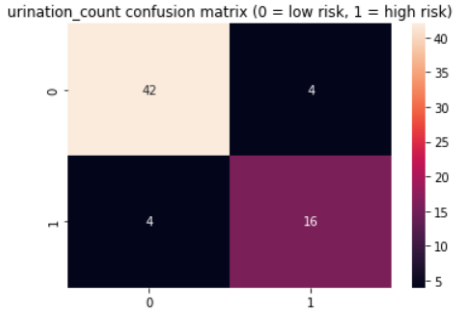


Figure A1: Selection process alpha and Lasso features

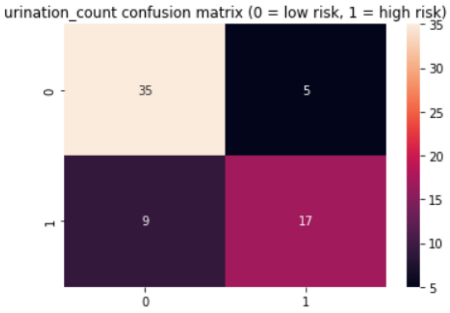
	precision	recall	f1-score	support
0	0.91	0.91	0.91	46
1	0.80	0.80	0.80	20
accuracy			0.88	66
macro avg	0.86	0.86	0.86	66
weighted avg	0.88	0.88	0.88	66

Wall time: 90.1 ms



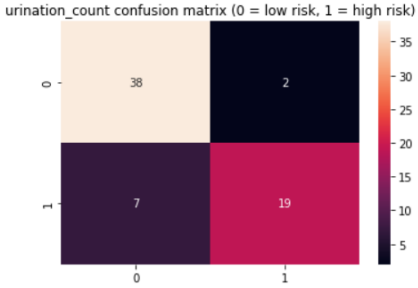
	precision	recall	f1-score	support
0	0.80	0.88	0.83	40
1	0.77	0.65	0.71	26
accuracy			0.79	66
macro avg	0.78	0.76	0.77	66
weighted avg	0.79	0.79	0.78	66

Wall time: 92.2 ms



	precision	recall	f1-score	support
0	0.84	0.95	0.89	40
1	0.90	0.73	0.81	26
accuracy			0.86	66
macro avg	0.87	0.84	0.85	66
weighted avg	0.87	0.86	0.86	66

Wall time: 324 ms



	precision	recall	f1-score	support
0	0.87	0.97	0.92	40
1	0.95	0.77	0.85	26
accuracy			0.89	66
macro avg	0.91	0.87	0.88	66
weighted avg	0.90	0.89	0.89	66

Wall time: 296 ms

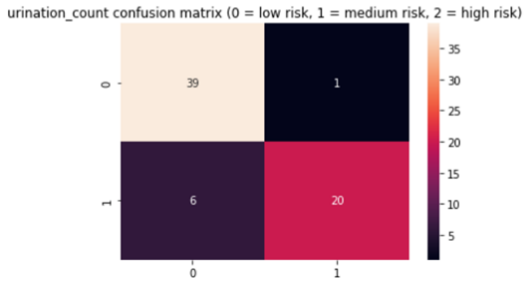


Figure A2: Confusion matrix of SVM (left) , DT (right), RF (bottom left), XGBoost

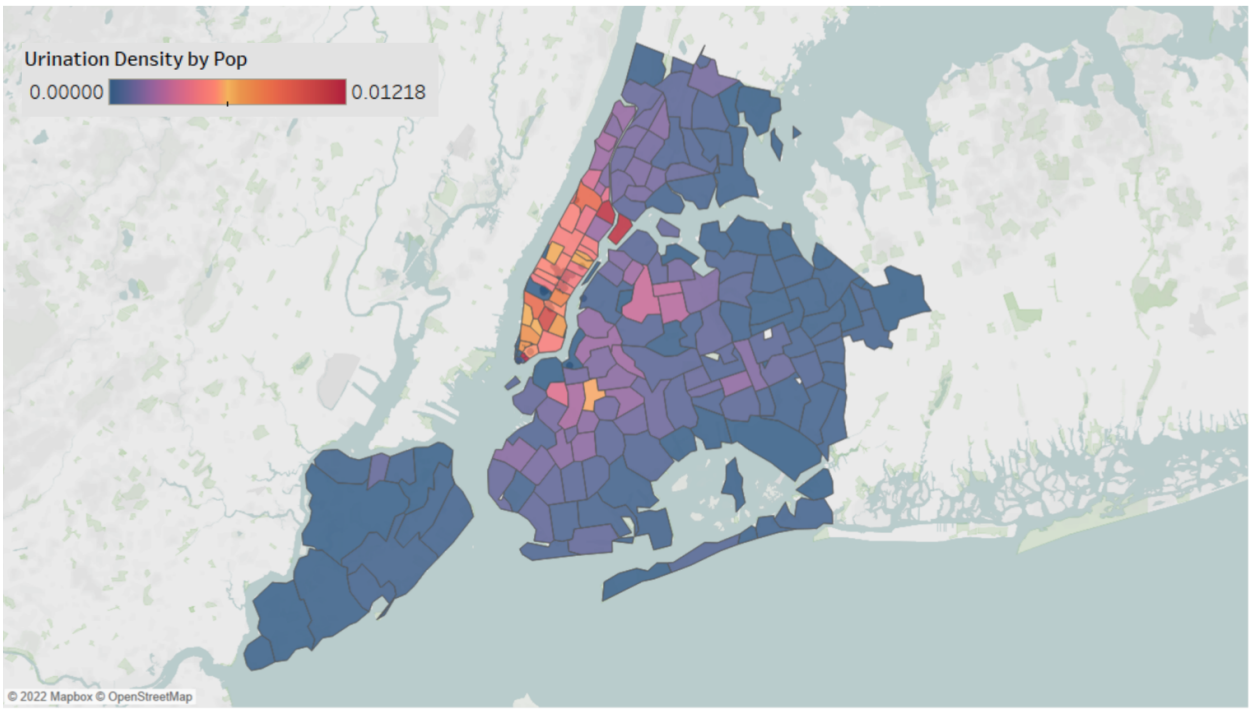


Figure A3: Exploration of urination density

Urination or defecation complaint density (by population) map (2010-2022)

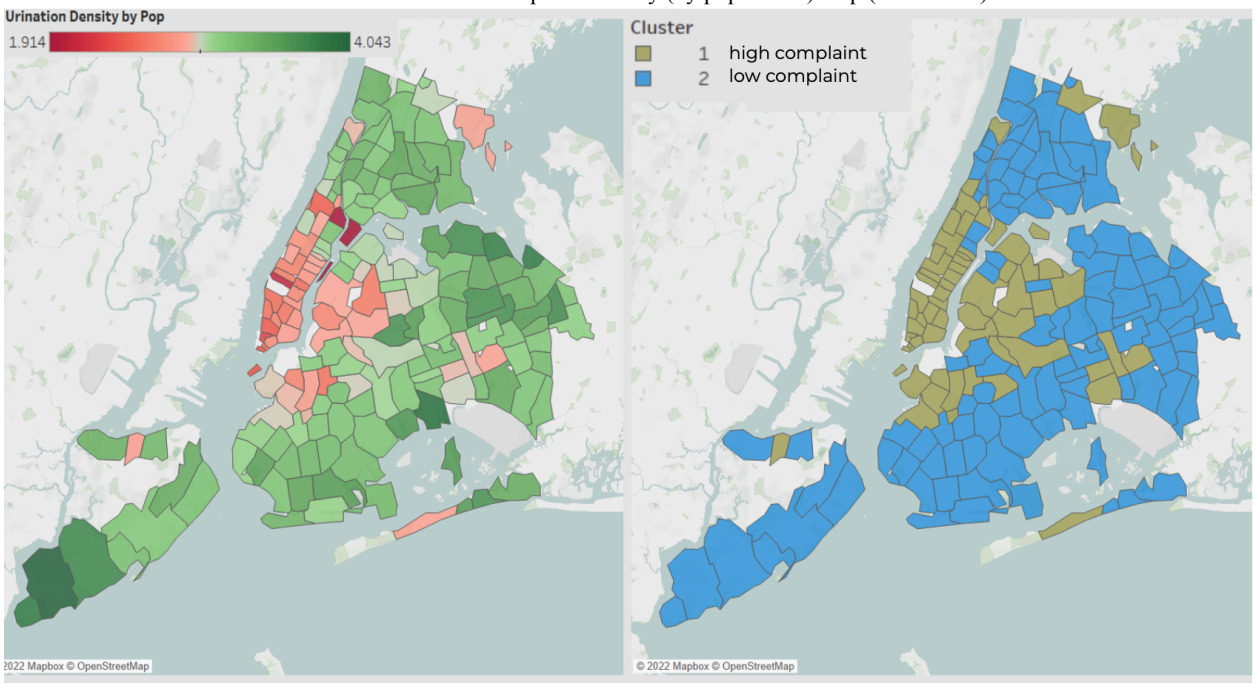


Figure A4: Urination or defecation complaint density (by population) map (2010-2022)

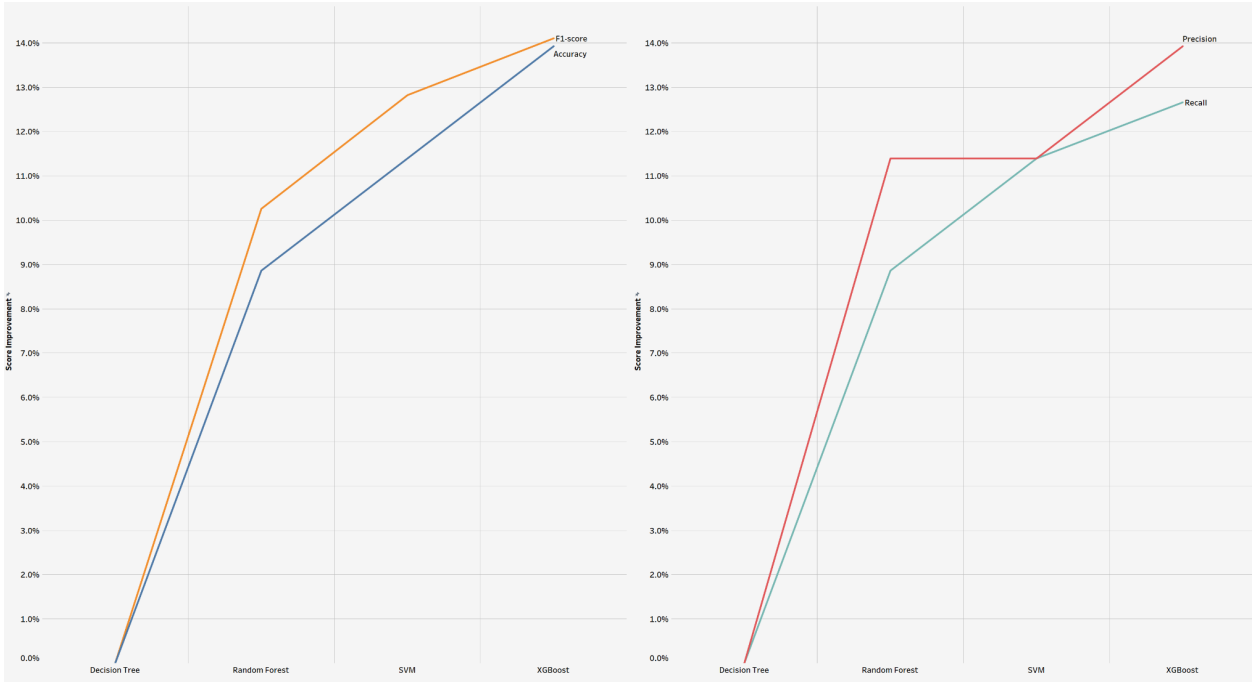


Figure A5:Percentage increase in model performance metrics (baseline: Decision Tree)

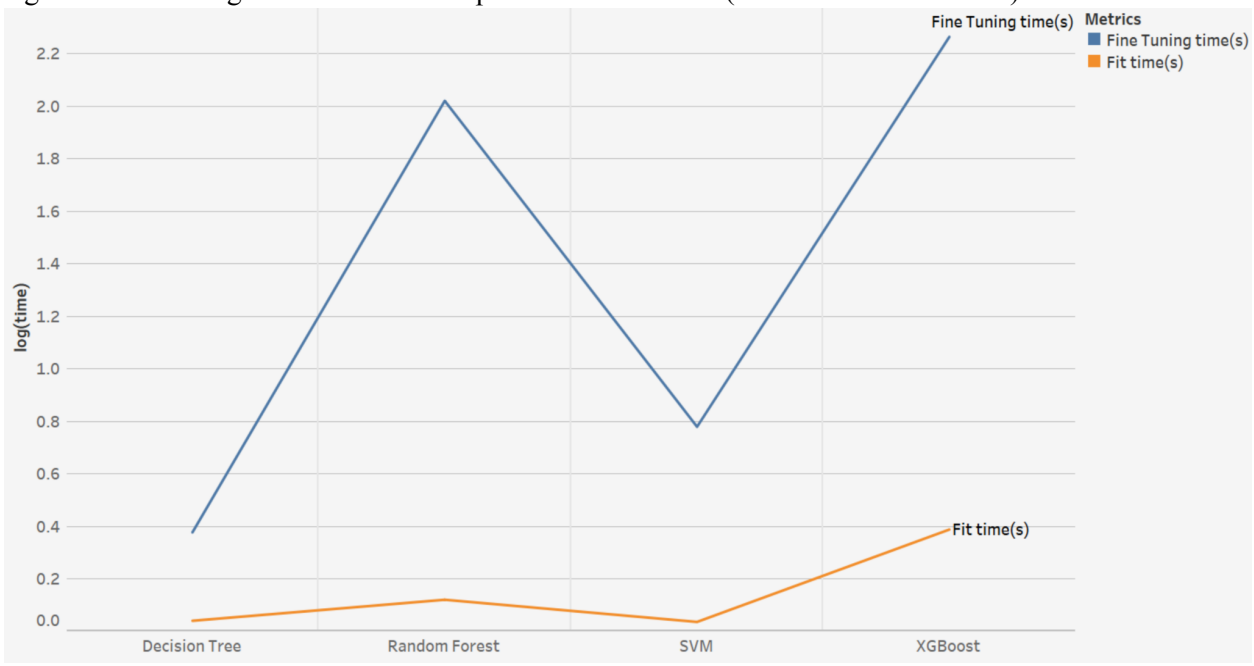


Figure A6: Model runtime summary

Appendix B (Table)

Table B1: All features used in model running.

Data		Description (Source)
Zip code		Zip code value of NYC (NYC Open Data)
Complaints (NYC Open Data)	Urination_count	Number of complaints in each zip code (311)
	Urination_density_pop	Density of complaints by zip code's population
	Urination_density_pop_degree	0: zip code with lower complaints 1: zip code with higher complaints
	Urination_density_area	Density of urination by zip code's area
Demographics for each zip code (ACS5)	PercentAgeUnder5	Percentage of population having age under 5
	PercentAge5to9	Percentage of population having age 5 to 9
	PercentAge10to14	Percentage of population having age 10 to 14
	PercentAge15to19	Percentage of population having age 15 to 19
	PercentAge20to24	Percentage of population having age 20 to 24
	PercentAge25to34	Percentage of population having age 25 to 34
	PercentAge35to44	Percentage of population having age 35 to 44
	PercentAge45to54	Percentage of population having age 45 to 54
	PercentAge55to59	Percentage of population having age 55 to 59
	PercentAge60to64	Percentage of population having age 60 to 64
	PercentAge65to74	Percentage of population having age 65 to 74
	PercentAge75to84	Percentage of population having age 75 to 84
	PercentAge85andOlder	Percentage of population having age over 85
	Male Percent	Percentage of Male
	Female Percent	Percentage of Female
	PercentWhite	Percentage of White people
	PercentBlack	Percentage of Black people
	PercentAmericanIndianandAlaska Native	Percentage of American Indian and Alaska Native

	PercentAsian	Percentage of Asian
	PercentNativeHawaiianandOtherPacificIslander	Percentage of Native Hawaiian and Other Pacific Islander
	PercentSomeOtherRace	Percentage of Some Other Race
	PercentHispanicOrLatino	Percentage of Hispanic Or Latino
Education Level (ACS5)	Less Than 9th Grade	Less Than 9th Grade
	9th to 12th Grade, No Diploma	9th to 12th Grade, No Diploma
	High School Graduate	High School Graduate
	Some College No Degree	Some College No Degree
	Associate Degree	Associate Degree
	Bachelor Degree	Bachelor Degree
	Graduate or Professional Degree	Graduate or Professional Degree
Income (ACS5)	PercentUnder$10,000	Percent of population with income under $10,000
	Percetn$10,000to$14,999	Percent of population with income from $10,000 to $14,999
	Percent$15,000to$24,999	Percent of population with income from $15,000 to $24,999
	Percetn$25,000to$34,999	Percent of population with income from $25,000 to $34,999
	Percent$35,000to$49,999	Percent of population with income from $35,000 to $44,999
	Percetn$50,000to$74,999	Percent of population with income from $50,000 to $74,999
	Percent$75,000to$99,999	Percent of population with income from $75,000 to $99,999
	Percetn$100,000to$149,999	Percent of population with income from $100,000 to $14,9999
	Percent$150,000to$199,999	Percent of population with income from $150,000 to $199,999
	Percetn$200,000ormore	Percent of population with income over $200,000
	Median Household Income	Median Household Income
Neighborhood Characteristics (NYC Open Data, OpenStreetMap)	tree_count	Number of trees in each zip code
	tree_density	Number of trees / area of zip code
	green_space_ratio	Area of green space / area of zip code
	shop_count	Number of shops in each zip code

	shop_density	Number of shops / area of zip code
	toilet_count	Number of public restroom in each zip code
	toilet_density	Number of public restroom / area of zip code
	crime_rate	Number of NYPD complaints / population of zip code
	population_density	Population / area of zip code
	subway_count	Number of subway entrances in each zip code
	AREA	Area of zip code
	POPULATION	Population of zip code