NYU

# Stinky NYC 🫠

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

Jingjing Ge | Yichen Guo | Jiashun Lian | Chaofan Zheng

Fall 2022 Machine Learning for Cities Final Project

December 2022

# Contents

- Introduction
- Related Works
- Our Approach
- Data Collection and Processing
- Methods and Algorithms
    - Exploratory variable selection with Lasso
    - Decision Tree, Random Forest, XGBoost
- Results
- Conclusion


Figure 1: Public urination in NYC (source: New York Post)

NYU

# Intro

## Before 2016:

- Public urination or defecation offense was misdemeanor under Public Health charge

## Since 2016:

- Public urination or defecation only results in civil penalties

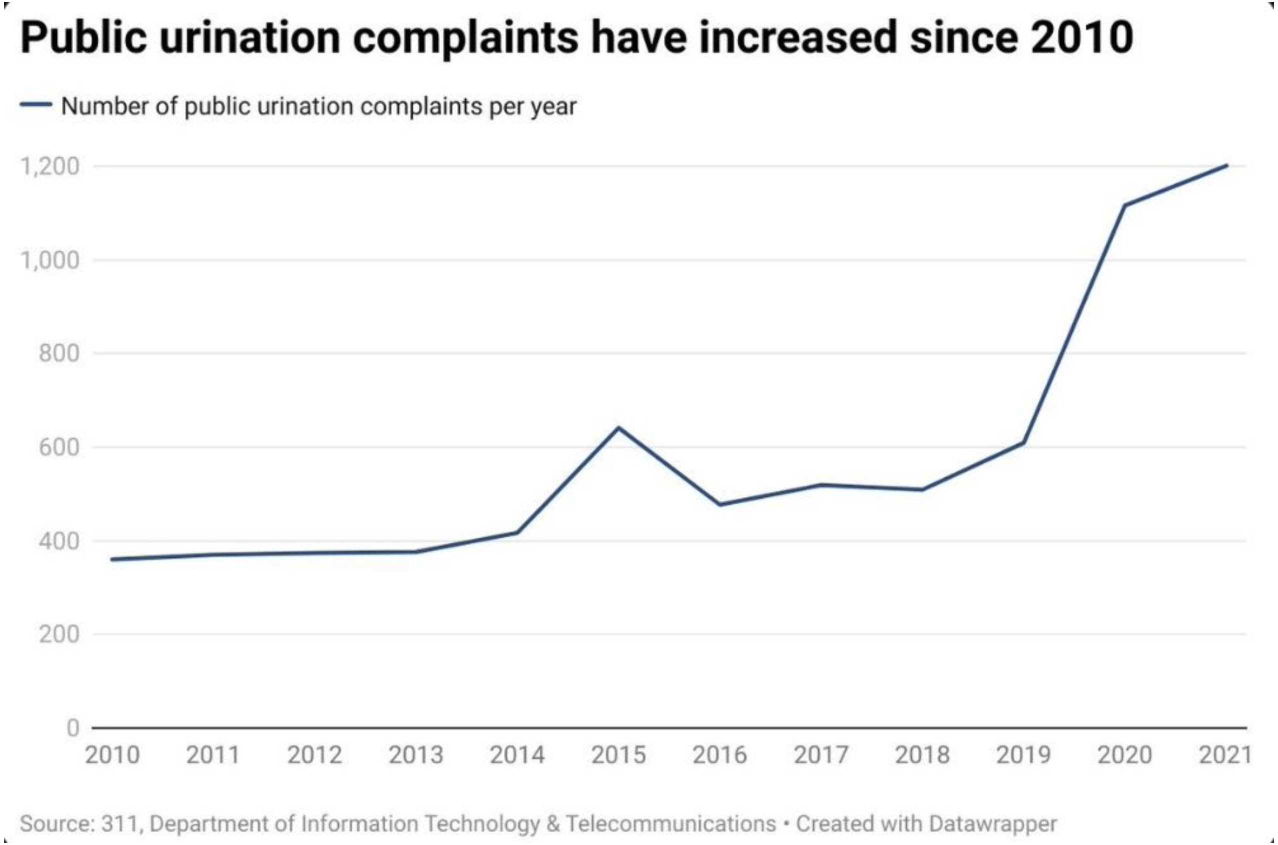



Figure 2: Public urination trend in NYC (source: Crain's New York)

NYU

# Related Works

- Machine learning approach on other complaints
  - Classification of crime (accuracy of 98%)
    - Logistic regression, multi-layer perceptron, decision tree and random forest
  - Alcohol-related complaints v.s. alcohol outlet density, area-level drinking, sociodemographic factors
    - Bayesian hierarchical Poisson regression
- Specifically on public urination or defecation
  - Only exploratory analyses were found

Figure 3: Something smells GIF (source: Gifer)

NYU

# Our Approach

Apply **machine learning** algorithms to target **NYC public urination and defecation**

- Classifying zip codes into high/low complaint area.
- Best model achieved
  - Accuracy: 90%
  - Precision and recall: 77% ~ 97%


Figure 4: Patrick smell float GIF (source: Tenor)

# Data Collection

- **Sources:** Publicly available
  - NYC Open Data
  - Census Bureau: API
  - Open street map: QGIS
- Main data: **Public Urination complaints**
  - 311 service requests from 2010 to present
- Geographical level: NYC **Zip code** boundary
  - Department of City Planning
- Features
  - Demographics, Public restroom, Green space, Commercial area .etc

Figure 5: Logos of Data sources

# Data processing

- Filter 311 “Complaint Type” to “urinating in public”
    - 7720 rows (year: 2010 - 2022)
- Spatial join using geopandas
    - Count number of complaints/feature values in each zip code
- **Normalized** by population or area based on feature type
- **Drop** any missing values
- Cleaned data
    - 167 rows: each represents a zip code
    - 57 columns: features

NYU

# Features Overview

- Zipcode
- urination_count
- urination_density_pop
- urination_density_pop_degree
- urination_density_area
- PrecentAgeUnder5
- PrecentAge5to9
- PrecentAge10to14
- PrecentAge15to19
- PrecentAge20to24
- PrecentAge25to34
- PrecentAge35to44
- PrecentAge45to54
- PrecentAge55to59
- PrecentAge60to64
- PrecentAge65to74
- PrecentAge75to84
- PrecentAge85andOlder
- Male Percent
- Female Percent

- PercentWhite
- PercetnBlack
- PercentAmericanIndianandAlaskaNative
- PercentAsian
- PercentNativeHawaiianandOtherPacificIslander
- PercentSomeOtherRace
- PercentHispanicOrLatino
- Less Than 9th Grade
- 9th to 12th Grade, No Diploma
- High School Graduate
- Some College No Degree
- Associate Degree
- Bachelor Degree
- Graduate or Professional Degree
- PercentUnder$10,000
- Percetn10, 000to14,999
- Percent15, 000to24,999
- Percetn25, 000to34,999
- Percent35, 000to49,999
- Percetn50, 000to74,999
- Percent75, 000to99,999
- Percetn100, 000to149,999
- Percent150, 000to199,999
- Percetn$200,000ormore

- tree_count
- tree_density
- green_space_ratio
- shop_count
- shop_density
- toilet_count
- toilet_density
- Median Household Income
- crime_rate
- population_density
- subway_count
- AREA
- POPULATION

NYU

# Data Exploration

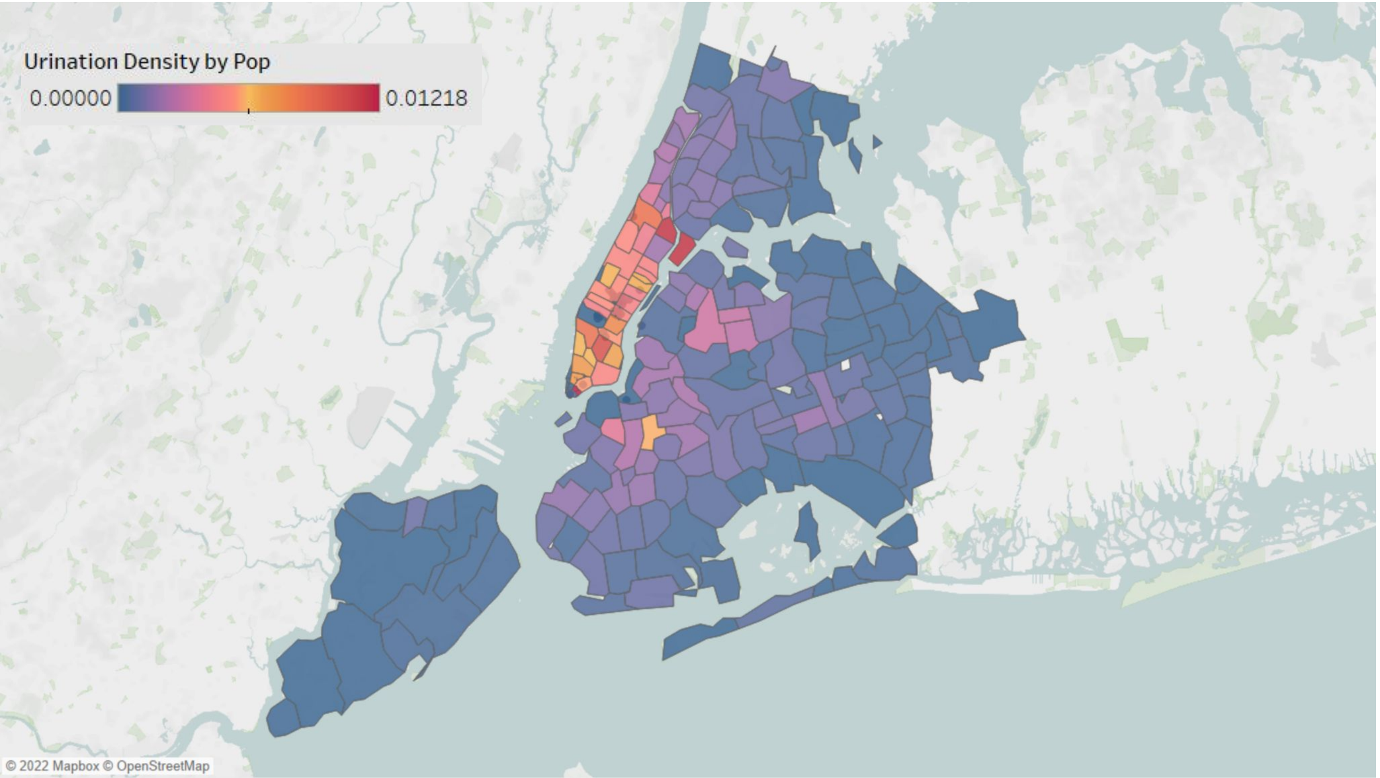



Figure 6: Urination or defecation complaint density (by population) map (2010-2022)

NYU

# Methods

**Linear Model (with features from Lasso):**

- A **simple and straightforward** way to model the relationship between a dependent variable and one or more independent variables.
- Easy to **interpret**
- Can provide a **measure** of the **strength of the relationship** between the predictor variables and the response.

NYU

# Methods

## Lasso Model(used for variable slection):

- **Reduce the number of variables**, useful for high-dimensional data. Because L1 regularization term make some variables' coefficients to become 0.
- Can also **eliminate multicollinearity** among variables, and **better interpret** the model.

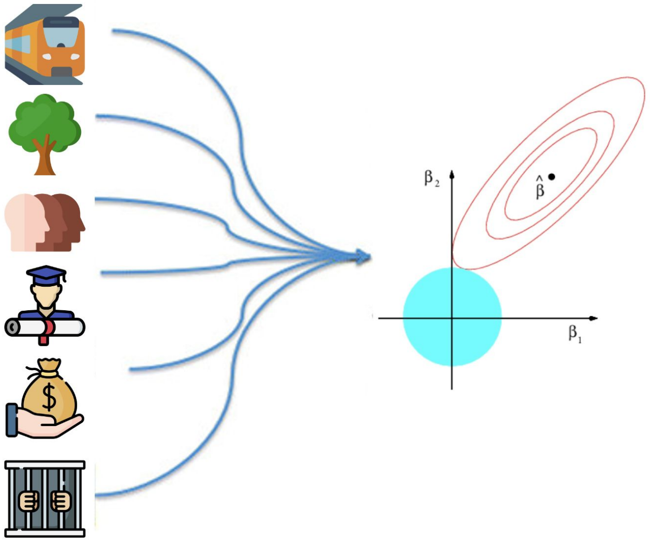


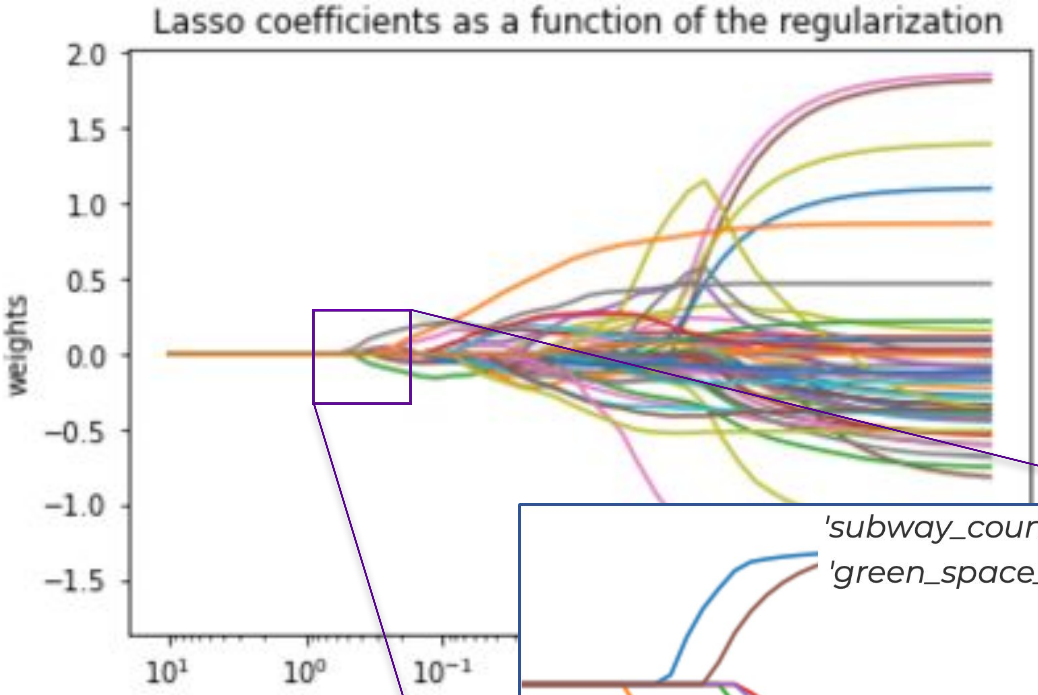

Figure 7: Lasso model chart

NYU

# Result

The **five variables** separated out are *'Associate_Degree','subway count','PercentAgeUnder 5',' highschoolGraduate','green_space_ratio'.*

Then, **multiple linear regression** is performed to predict urination density.



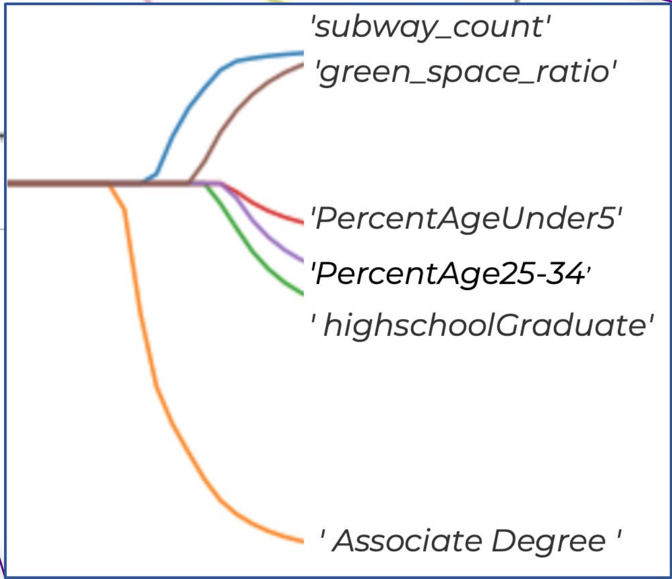




Figure 8: Urination or defecation complaint density (by population) map (2010-2022)

NYU

```
                            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 9: OLS regression result

After **avoiding a high degree of covariance**, the in-sample model fit is not ideal, with an **R squared of 0.480**. Clearly, this indicates that they are **not the main influencing features** in the model. And **out of sample accuracy is 0.446140**. this clearly did not achieve the expected results. Therefore, in summary, the multivariate linear model is poorly fitted.

NYU

# Classification Model

Due to the significant differences between regions, we believe that a classification model is better to handle this problem

# How do we divide the New York region into different levels?

*“K-Means sounds good.”*

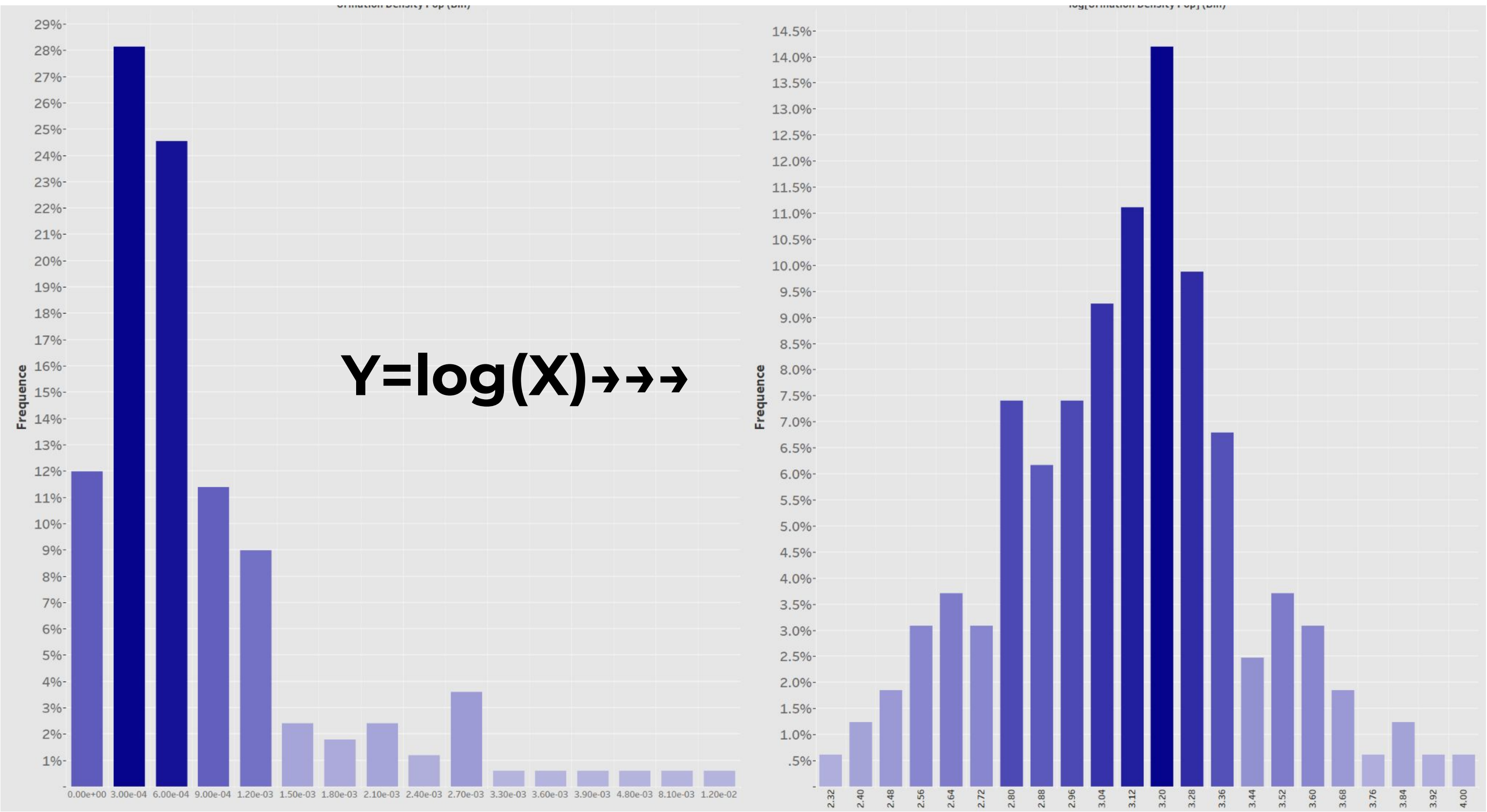



Figure 10: distribution of Urination or defecation complaint density (by population) map (2010-2022)

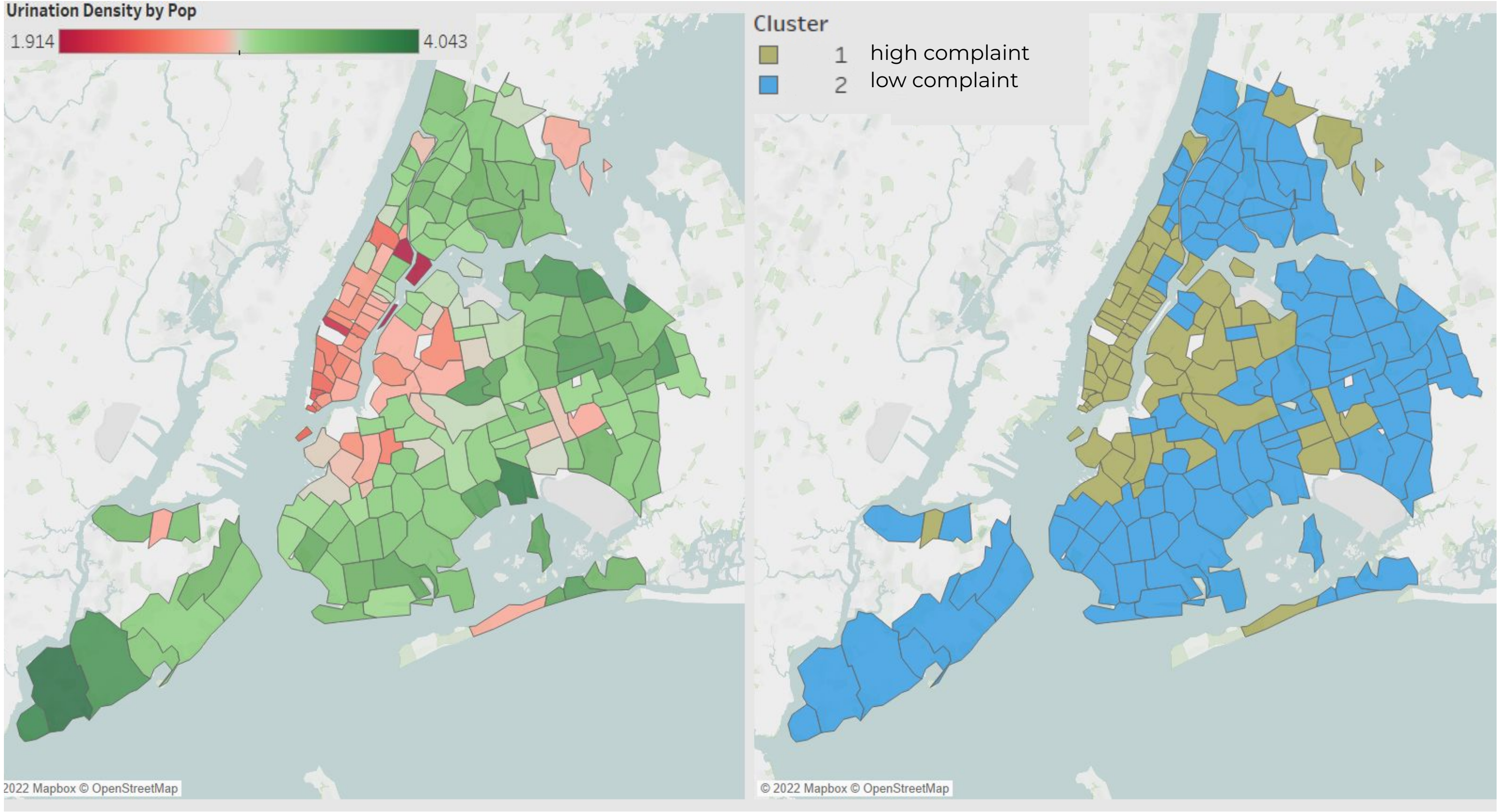



Figure 11: Urination or defecation complaint density (by population) map (2010-2022)

NYU

# Method (SVM)

SVM is kind of model that can be effective in high dimensional spaces. By using a **kernel trick** to transform the data into a **higher-dimensional space** where it can be **linearly separated**. This makes SVMs a useful tool for a variety of applications.

| | 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

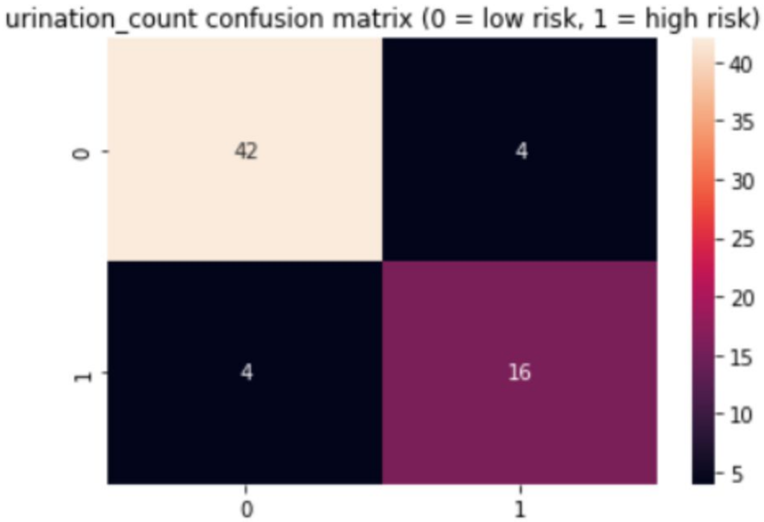



Figure #: Support vector machine performance summary

# Method (Tree Model)

Tree models have many advantages.

- Easy to **understand** and **interpret**
- Handle **unrelated features** and mine **non-linear relationships**
- Handle both **numerical and categorical** features.

NYU

# Method (Tree Model)

We selected three representative models from tree models.

| | |
|---|---|
| Decision Tree | The most basic tree model. |
| Random Forest | An ensemble learning method that constructs multiple decision trees. |
| XGBoost | An implementation of gradient boosted decision trees. |

| | 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

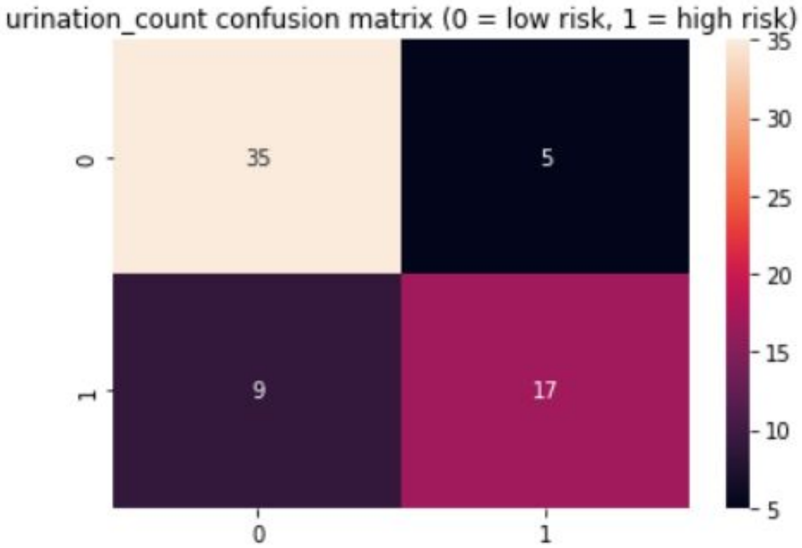



Figure #: Decision tree performance summary

| | 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

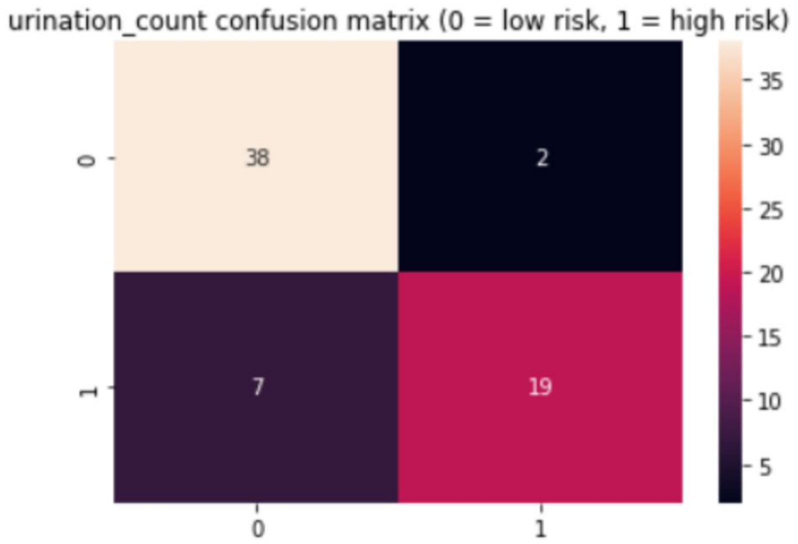



Figure #: Random forest performance summary

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.97 | 0.92 | 31 |
| 1 | 0.94 | 0.79 | 0.86 | 19 |
| accuracy | | | 0.90 | 50 |
| macro avg | 0.91 | 0.88 | 0.89 | 50 |
| weighted avg | 0.90 | 0.90 | 0.90 | 50 |

Wall time: 1.1 s

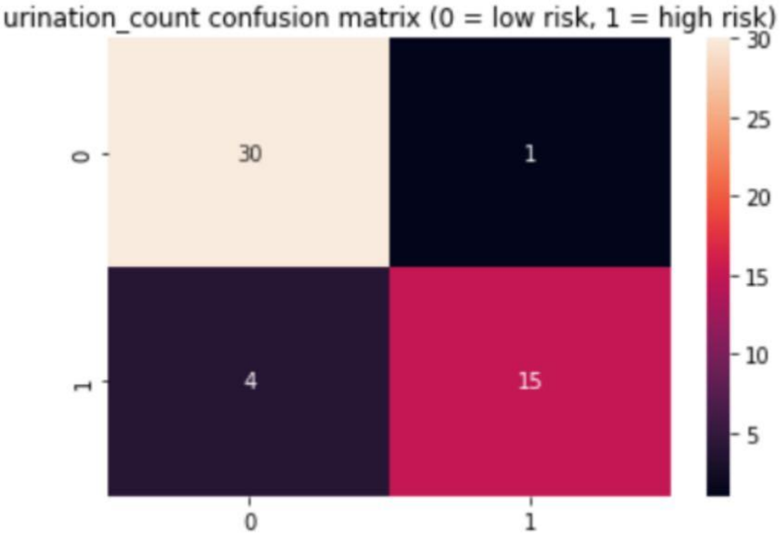



Figure #: XGBoost performance summary

| 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 12: Model performance summary

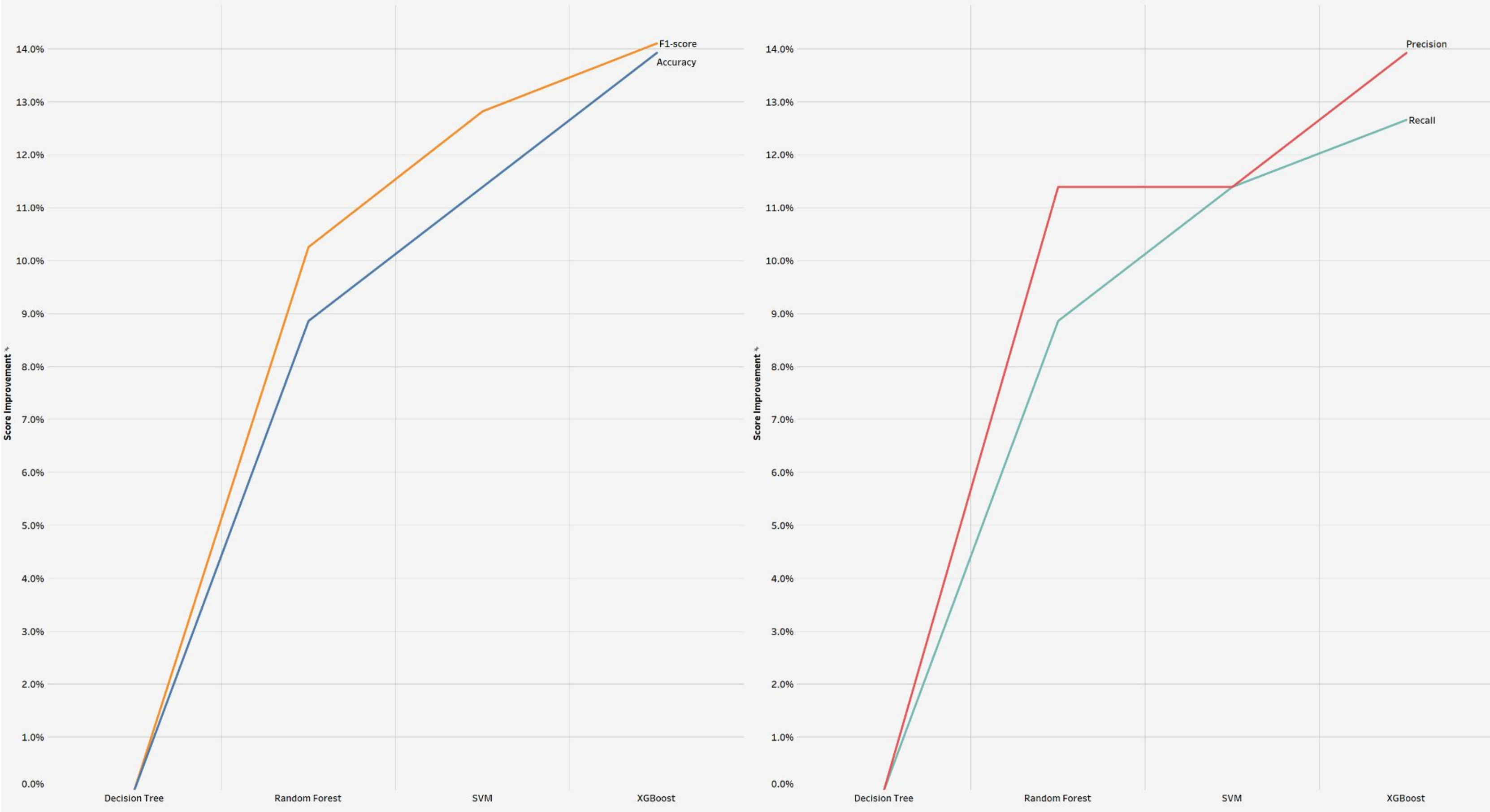



Figure 13: Percentage increase in model performance metrics (baseline: Decision Tree)

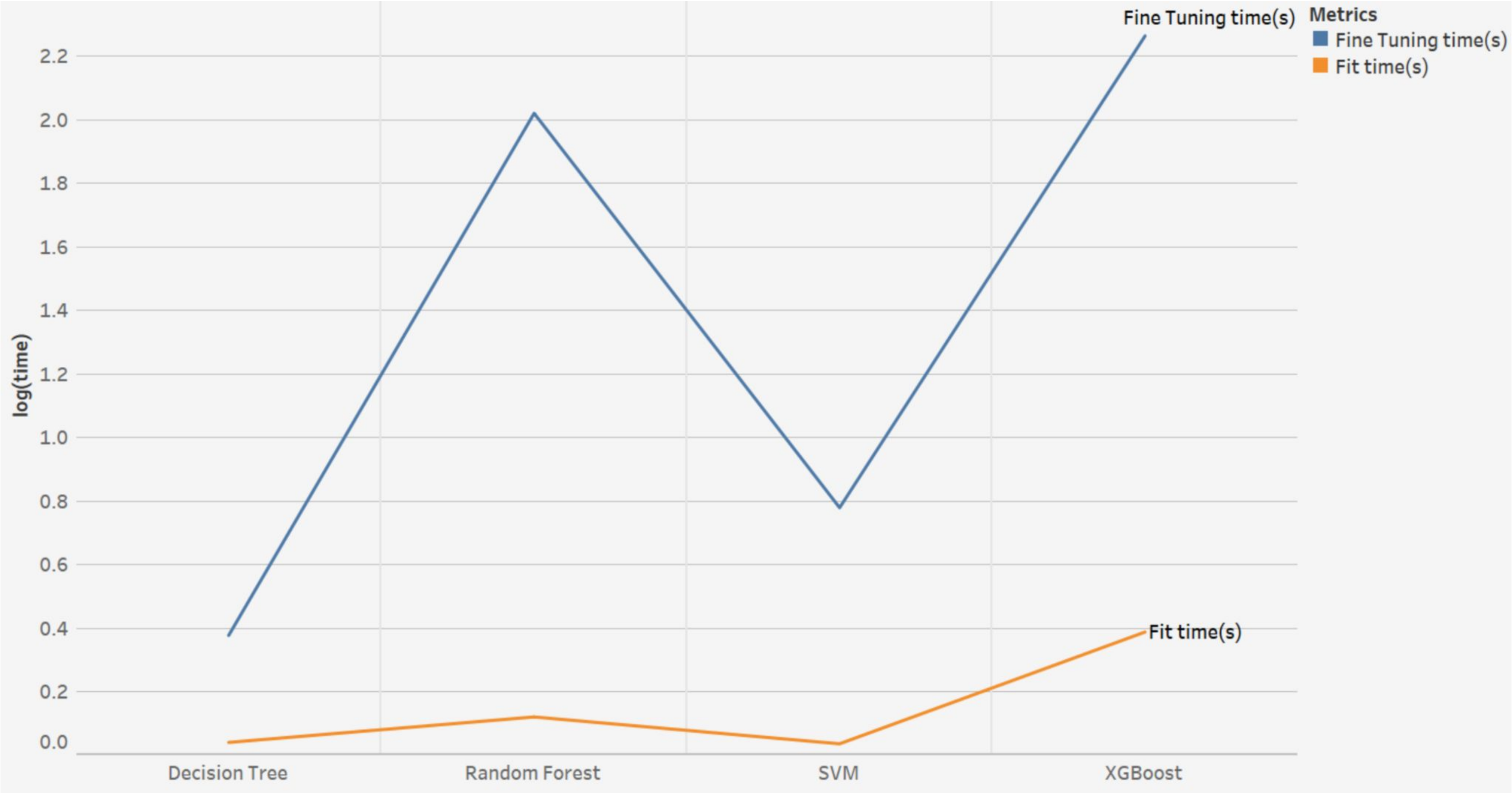



Figure 14: Model runtime summary

# SHAP Explainer

*“**SHAP (SHapley Additive exPlanations)** is a game theoretic approach to 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”*

*– SHAP Developer*

—> Increase **transparency** and **interpretability** of our models

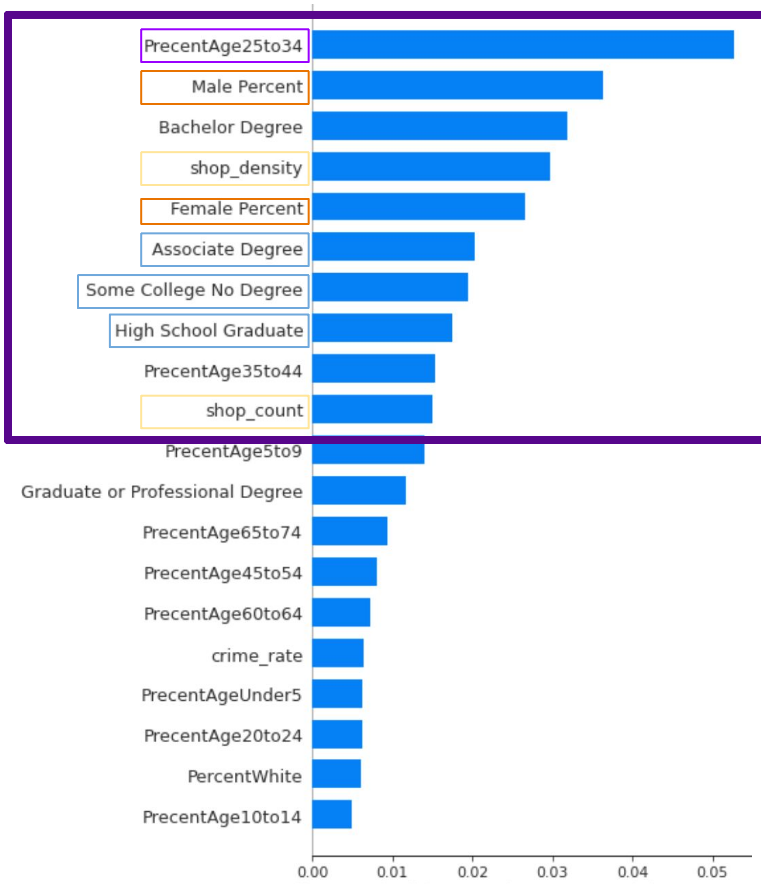



Figure 15: Random forest mean absolute SHAP value

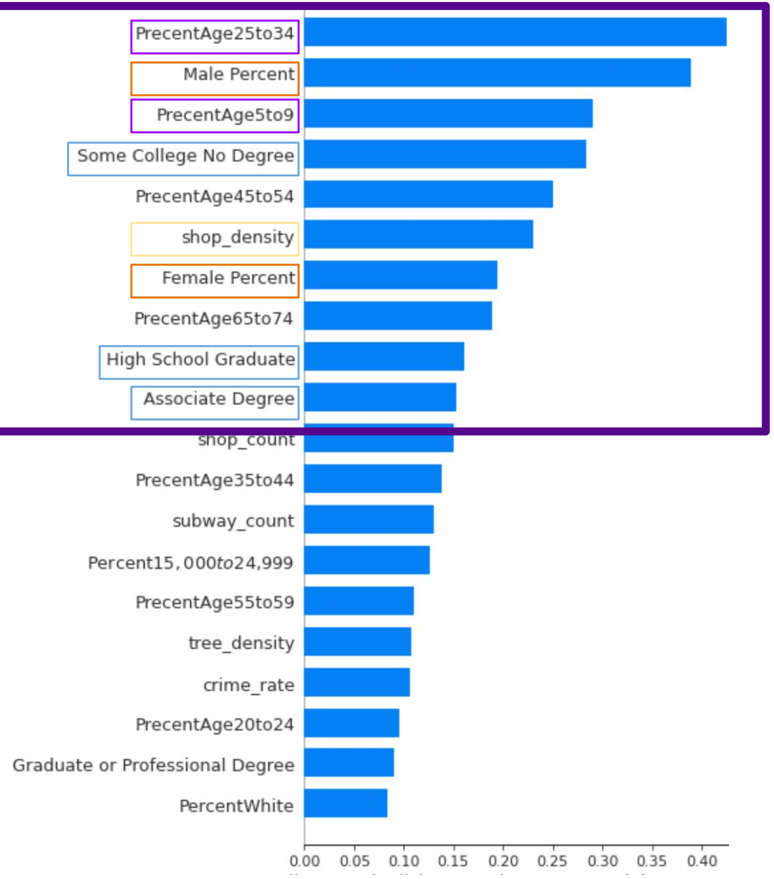



Figure 16: XGBoost mean SHAP value

NYU

**Represent education level** **Represent gender** **Represent age group** **Represent shop**

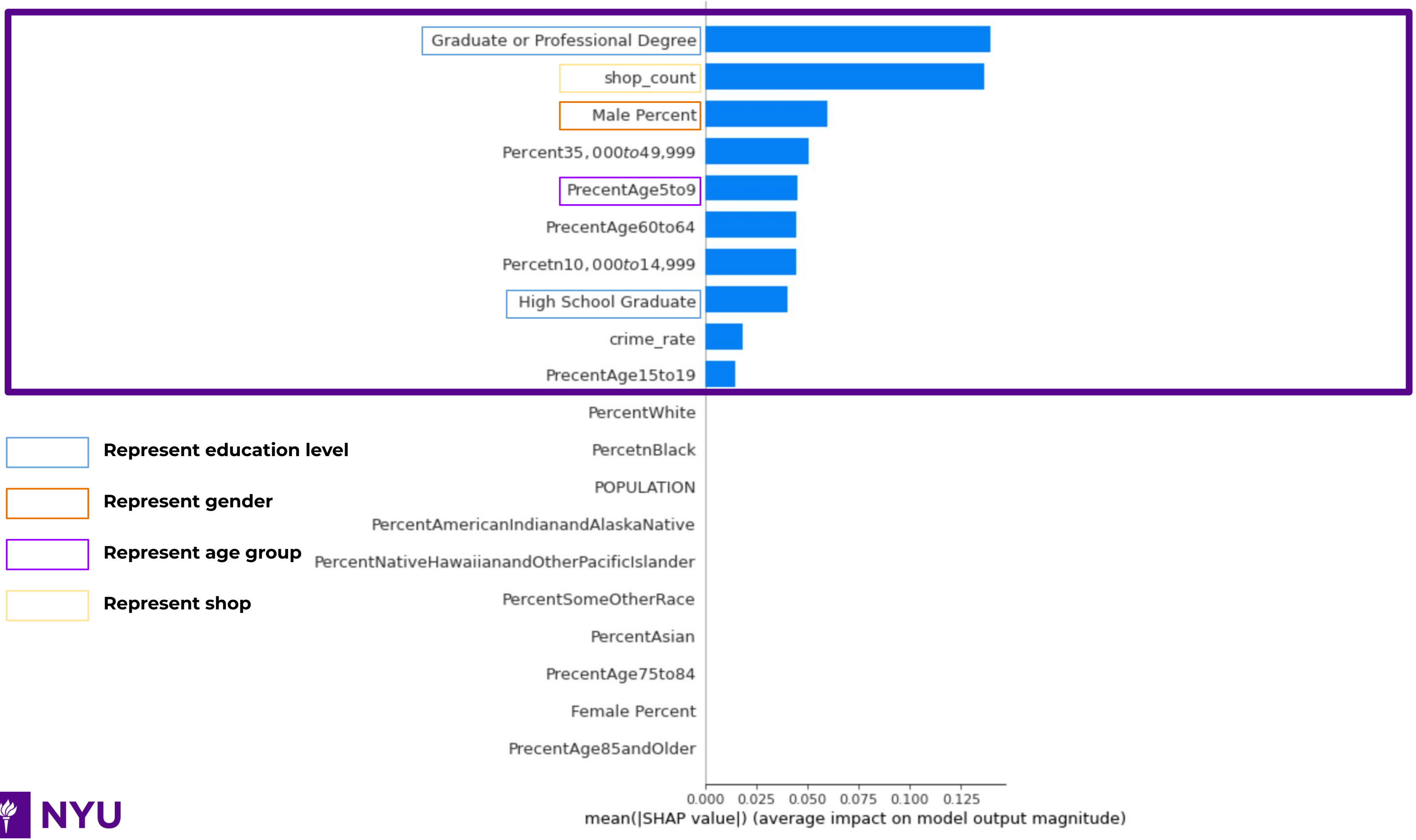



Figure 17: Decision tree mean SHAP value

Represent education level  Represent gender  Represent age group  Represent shop

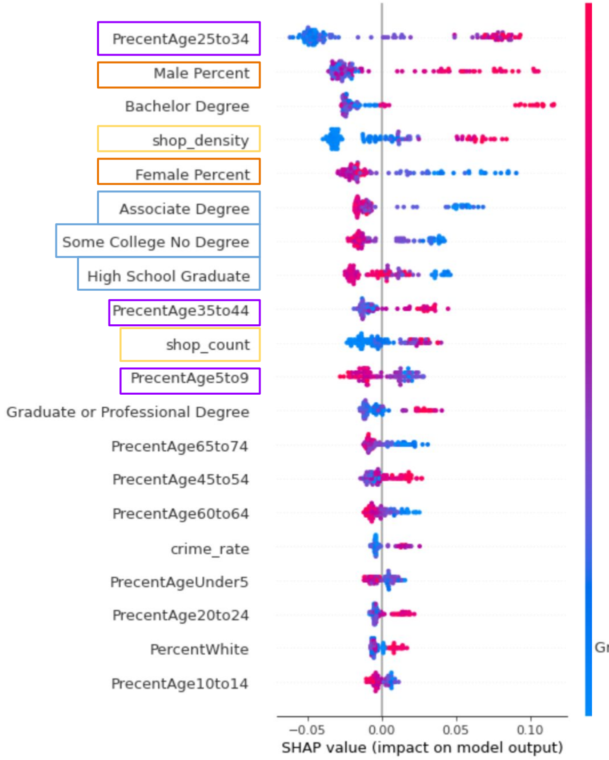



Figure 18: Random forest SHAP summary

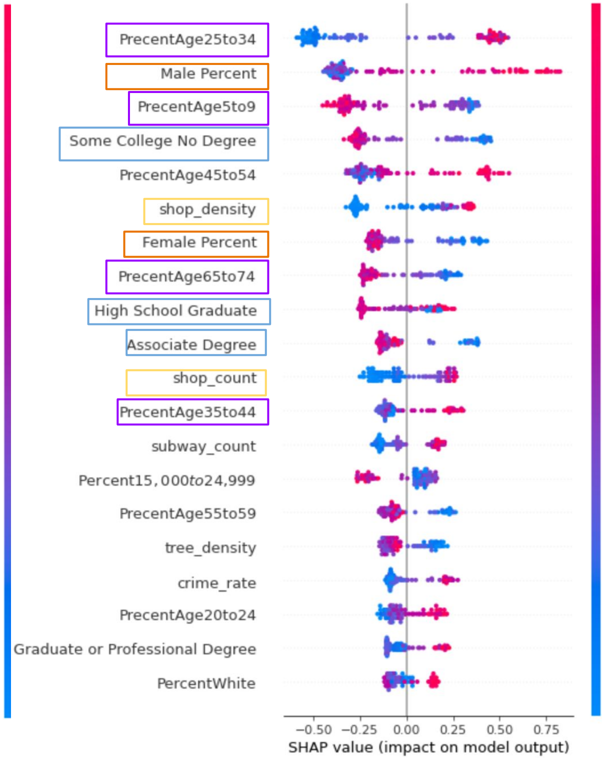



Figure 19: XGBoost SHAP summary

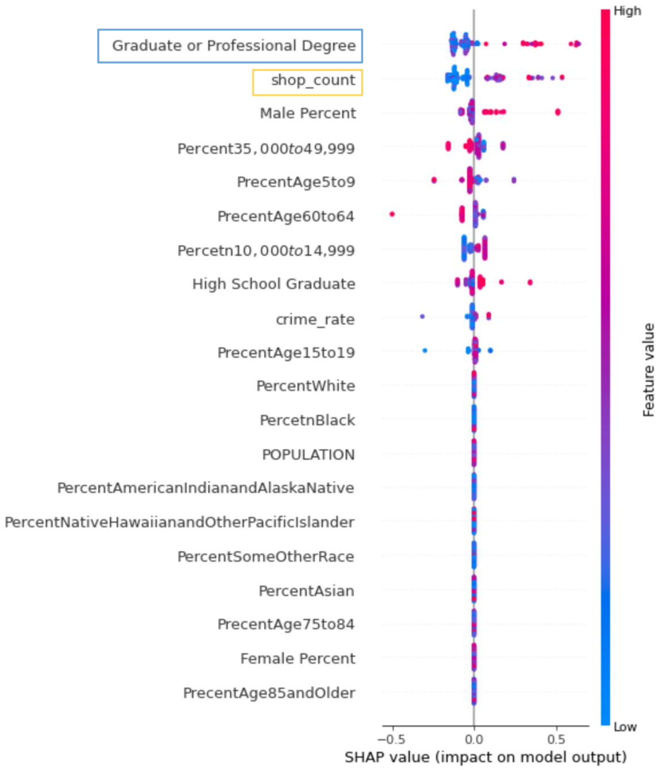



Figure 20: Decision tree SHAP summary

NYU

# Conclusion

- Lasso for feature selection
  - **Positive** correlation
    - **subway count** & **green space%**
  - **Negative** correlation
    - **associate degree%** & **high school graduate%** & **age under 5%**
- Linear regression
  - Lack performance in this scenario

# Conclusion

**RF, XGBoost, DT (SHAP)**:

- Most correlated: **age, gender, education level, shop**
- Relatively correlated: **income level (only in DT)**

↓ complaint is associated with:

- ↓ **age 25-34 %, male%, shop%**
- ↑ **high school graduate%, associate degree%, some college% (education)**
- ↑ **age 5-9%, age 65-74%, female%**

NYU

# Insights

| "GOOD" Areas: | | "BAD" Areas: |
|---|---|---|
| Well-educated (at least high school %) - | ≠ | - Less educated ? |
| Less people from 25-34 - | | - More people from 25-34 |
| Less shops - | ≠ | - More shops ? |
| Less males - | ≠ | - More males ? |
| More people from 5-9 - | | - Less people from 5-9 |
| More people from 65-74 - | | - More people from 65-74 |

*Younger people from **25-34** tend to **file more complaints**, while **elders not familiar** with **311** for complaints*

*Parents with kids from **age 5-9** tend to choose **better living environments***

***Less shop** areas have **less passengers**, thus **lower complaints***

***Well educated** areas have **less occurrences***

NYU

# Future Steps

- **Smaller** geographical level
- **COVID** specific trend
- **Seasonal trend** analysis (weather, temperature, time)
- Explore **other odor** complaint (chemicals, trash, etc)
- **Oversampling** to expand our dataset

# Thank you for listening

THE END

NYU

# Reference

https://nypost.com/2022/07/14/nyc-odor-complaints-reach-disgusting-all-time-high/
https://www.kaggle.com/code/jasonduncanwilson/urination-in-nyc-and-other-fun-exploration/notebook
https://aidalalaw.com/is-new-york-city-about-to-ease-up-on-petty-crimes/
https://www.pinksummons.com/public-urination-health-code-summons-nyc
https://github.com/htappa/NYC_CrimeData/blob/master/NYPD_CrimeData.ipynb
https://www.foxnews.com/us/new-york-city-eases-severity-of-laws-against-public-urination-drunkenness
https://twitter.com/CrainsNewYork/status/1496189259946106882
https://gifer.com/fr/gifs/something-smells
https://tenor.com/view/patrick-smell-float-bon-sponge-gif-15840520
https://doi.org/10.1007/s11524-018-00327-z

# 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