

cs109a Final Project, Milestone 3
Project C: Predicting Types of Crime
Project team #109

Contributors:
Christopher Campion
Sheraz Choudhary
Fabio Pruneri
Michael Sedelmeyer

Contents

Description of the data	2
Data Sources Used	2
Noteworthy EDA findings	2
Crime Incident Data	2
Property Assessment Data	4
Streetlight Location Data	5
Neighborhood Demographics Data	5
Education Institutions	6
Property Violations	6
Liquor License Data	7
Revised project questions	8
Baseline model	8
Response Variable	8
Predictors	9
Model Results	10
Appendices	10
Appendix 1: Predictors still under development	10
Appendix 2: Overview of predictor collinearity	11
Appendix 3: Estimated LASSO regularized multinomial logistic regression model coefficients by response class	13

PLEASE NOTE: Besides any sample notebooks submitted with this milestone, all project work completed to-date by our team (including all notebooks in progress and supporting materials) can be viewed directly on GitHub in our shared project repository located here: <https://github.com/sedelmeyer/predicting-crime>

Description of the data

Data Sources Used

Below is a list of [Boston.gov](https://www.boston.gov) data sources investigated as part of our initial EDA and used to assist in feature engineering for our baseline model. For a complete list of URLs for the below listed datasets, and to see all other datasets we have found and are investigating to-date (including data from additional sources), please see the “[data-inventory.csv](#)” file in our GitHub repository.

1. Crime incident data
2. Property assessment data
3. Streetlight location data
4. Street Address Management (SAM) system data
5. Neighborhood demographics data
6. Liquor licensing data
7. Public and non-public schools data
8. Universities and colleges data
9. Various City of Boston shape files
 - These include Census tracts, Boston neighborhoods, Zip codes, Street segments, and Open spaces

Noteworthy EDA findings

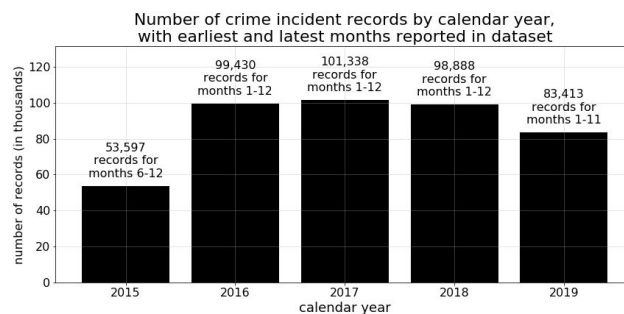
Please note that many additional charts and EDA findings can be found in the accompanying Jupyter notebooks, which can also all be found on our [project GitHub](#) at the links accompanying each sub-section below.

Crime Incident Data

EDA Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/009_EDA_crime_incident_reports.ipynb

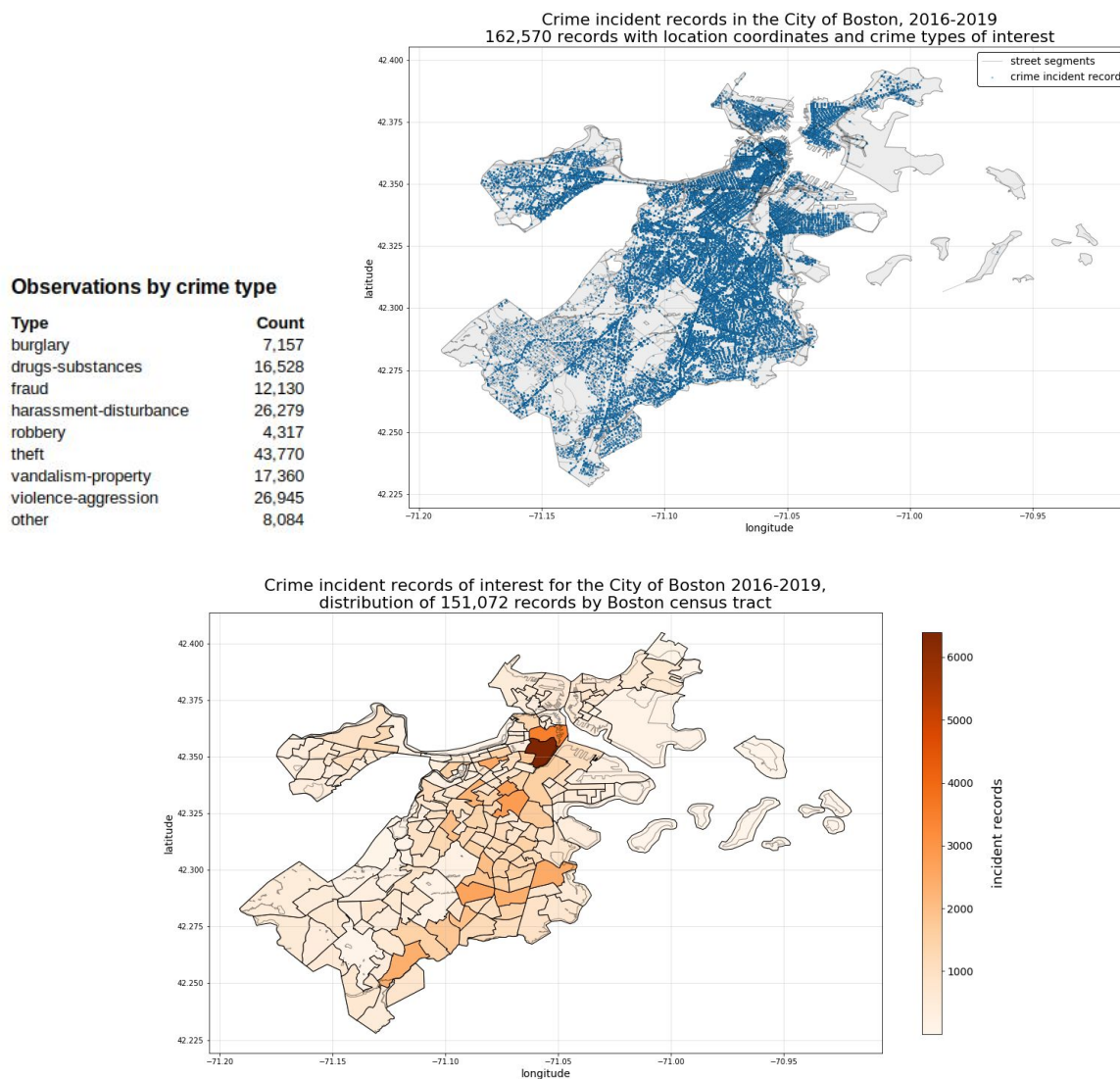
The City of Boston’s crime incident reports (August 2015 - to date) data is the source data for the parameter of interest of our analysis, crime type. The [original data](#) contained 436,666 crime incident observations, spanning 66 different crime “offense code groups” across the City of Boston.



Ultimately, due to limitations imposed by years of available property assessment data (see next section), and the limited number of 2015 observations, we have decided to use incident reports for just the 2016-2019 calendar years. Because of the geospatial nature of the analysis we have undertaken, we also needed to exclude 35,785 additional records with missing Latitude and Longitude coordinates. Due to a disproportionately

large proportion of records during the final 3 months of observations (Sep.-Nov 2019) missing location coordinates, those three months were also excluded in their entirety.

Following the subsetting of our data from Jan. 2016 through Aug. 2019, and excluding our observations with missing coordinates, we were left with 347,284 crime incident observations for building our analysis. Of these remaining records, because we are ultimately dealing with a classification challenge, we felt it was important to consolidate our 66 offense code groups into a smaller subset of “crime types” to generate more meaningful results later in our analysis by (a) removing incident categories of little importance that might otherwise obscure trends in more Important areas of crime (for instance removing things like ambiguous “investigations” or non-crime incidents like “motor vehicle accident response”) and (b) consolidating the remaining categories into a set of 9 different “crime types” (burglary, drugs-substances, fraud, harassment-disturbance, robbery, theft, vandalism-property, violence-aggression, other). Then, once we drop unused offense codes and are able to tie each of record to its corresponding census tract shape (the geospatial unit of analysis we ultimately use for summarizing property-related features), we are left with 151,072 records remaining for our analysis, distributed as shown below:



Please note that additional analyses were conducted as part of this EDA, but due to limited space, cannot be included here. By referring to [our accompanying notebook](#), you can also view our analysis to investigate

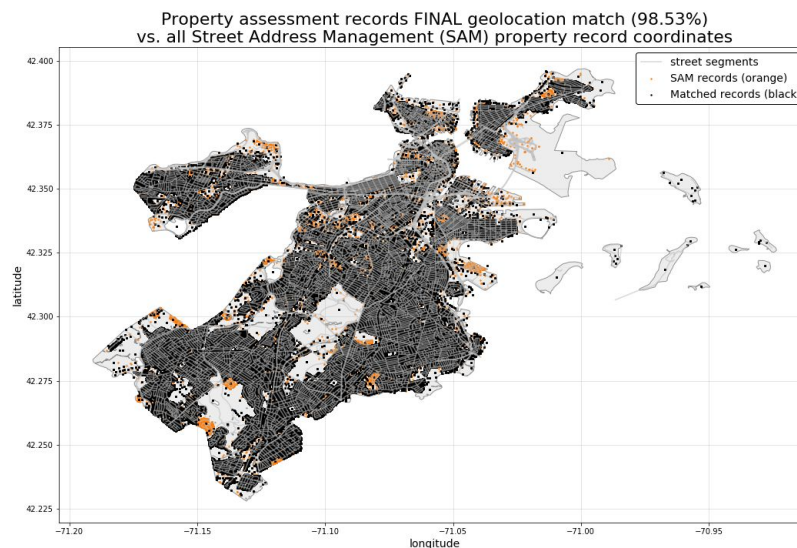
potential observation bias in the days, months, hours, and crime type categories of each observation with location coordinates versus the observations with missing coordinates.

Property Assessment Data

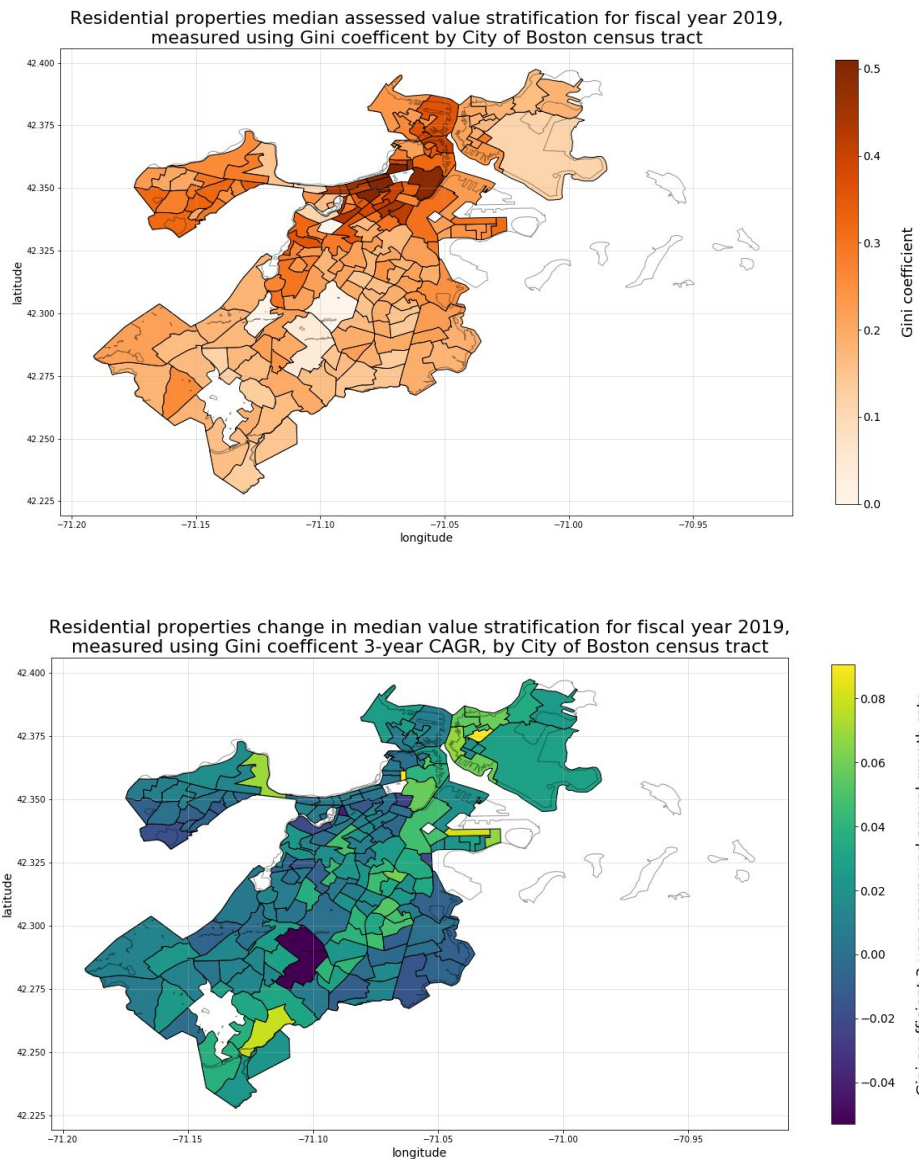
EDA Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/010_EDA_boston_property_assessments.ipynb

The City of Boston's property assessment records provided data ultimately used to engineer several features for our baseline model. The records used for our analysis can be found on the [Boston Analyze website](#) and included all Boston properties (by Parcel ID [PID] and address) from fiscal years 2013 to 2019. The overall set of data investigated included 1,185,432 property records across all years and the largest challenge encountered working with this data was matching Latitude and Longitude coordinates to years of data that did not include coordinates with their respective data files (approx. 1,083,590 records). Based on PID for years reporting coordinates, the vast majority of these were matched. For the remaining 78,018 unmatched records, the City of Boston's Live Street Address Management (SAM) System [parcel dataset](#) was used for matching coordinates. Using this SAM dataset, we were able to match coordinates to all but 32,297 assessment records (98.5% total match rate) as is summarized in the plot below:



Then, with these matched coordinates, we were able to explore the remaining dataset for features of interest as well as decide on geospatial shapes for use in defining a set of baseline model predictors based on our remaining property assessment records. We ultimately engineered a set of 10 property-based predictors, each providing a census tract-level metric per-year to give a point in time measure for that geographic area, as well as a 3-year average annual change rate, to measure the magnitude of shifting property demographics for each area. The full set of features are described below in the "Baseline Model" section of this report. However, to illustrate a sampling of these engineered features, see the two plots below:



Once again, for a more complete overview of all the EDA, data cleansing, and feature engineering methods used for the property assessment dataset, [please refer to the accompanying notebook](#).

Streetlight Location Data

SHERAZ

Neighborhood Demographics Data

EDA Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/015_EDA_neighborhood_demographic_s.ipynb

By Boston neighborhood we have successfully pulled together demographic data in the following categories:

- Age
- Housing Tenure

- Household Income
- Poverty Rate
- Educational Attainment | School Enrollment

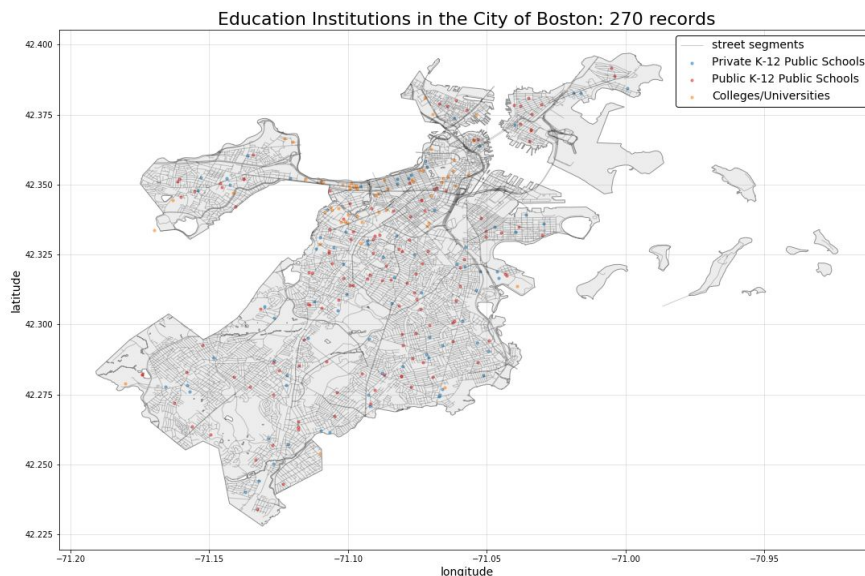
For our 17 identified neighborhoods we have created one master data frame including over 70 features granting us flexibility for model experimentation. Given that the demographics data is at an aggregated neighborhood level (rather than by census tract) further discussion is needed around model implementation.

Education Institutions

EDA Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/014_EDA_schools_and_universities.ipynb

We have examined public K-12 schools, non-public K-12 schools, as well as colleges and universities in the Boston area. We've cleaned and prepared 131 Public K-12 Boston area schools, 82 non-public schools, and 57 colleges/universities. The plot below illustrates the combined 270 education institutions differentiated by color. All education institutions include geographical coordinates and are ready to be utilized within our models.



Property Violations

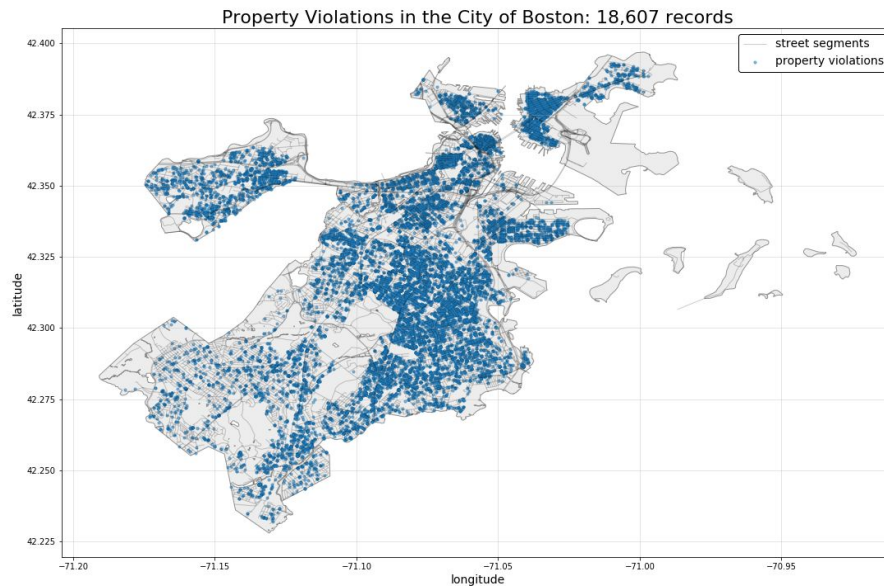
EDA Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/016_EDA_property_violations.ipynb

Over 18,000 property violations in the Boston area have been analyzed across 453 property violation types. The most common property violation types include:

1. "Unsafe and Dangerous" (17%)
2. "Failure to Obtain Permit" (14%)
3. "Owners Responsibility to Maintain Structural Elements" (8%)

Given that the remaining 450 property violations cover the remaining 69% of all violations we may consider further feature engineering to simplify our violation types or exclude property violations from our models. The plot below illustrates all property violations in the city of Boston.



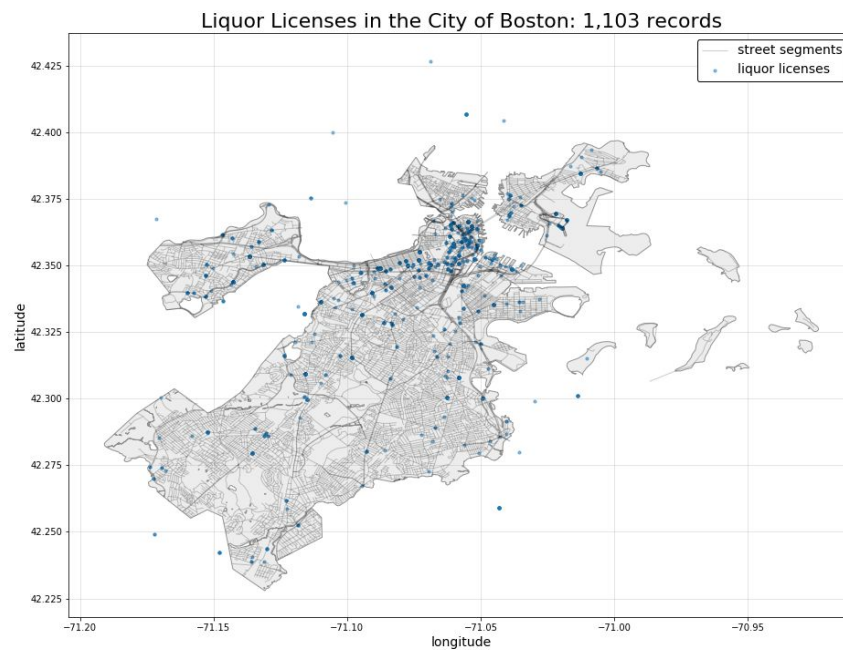
Liquor License Data

EDA Notebooks:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/012_liquor_data_prep.ipynb

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/013_EDA_liquor_data.ipynb

Examining liquor licenses in Boston took a fair amount of data cleansing efforts to transform incomplete addresses to latitude and longitude. We have successfully processed over 1,100 liquor license records and plotted them in the figure below:



What we've found a bit peculiar is that several of our data points fall outside our map. Digging a bit deeper we've found that over 85% of our liquor license categories belong to "Common Victualler" which represents "any establishment that has on its premises the ability to assemble, prepare, or cook food". We would not

expect such an unequal distribution of liquor license types which raises concern towards the completeness of our dataset. Given these concerns we may decide to exclude the presence of liquor licenses in our final model.

Revised project questions

Given the (1) variety preliminary datasets we were able to find, (2) the results of our initial EDA on a subset of those datasets, and (3) the number of additional predictors we currently have under development (see the Next Steps section of this document), we have decided to define our project questions in fairly broad terms similar to how they were proposed in the original project description :

- Given a specific location in the City of Boston, can we predict (from a defined subset of crime types), which type of crime is most likely to occur at that location?
- What effects do location-based features have on the type of crime most likely to occur at that location?

Baseline model

Baseline Model Notebook:

https://github.com/sedelmeyer/predicting-crime/blob/master/notebooks/021_MODEL_baseline_logistic_classifier.ipynb

As an initial baseline model, we ran several multi-class Logistic Regression models on a version of our predictors outlined below, in which all non-binary predictors were standardized to adjust for large variability in scale among predictors. Variations attempted while building our baseline model included both one-vs-rest and multinomial versions of the model. In addition, we ran the versions of the models without regularization and then with L1 Lasso-like regularization (but without cross-validation) to ultimately examine coefficient shrinkage and to begin understanding relationships between our response classes and each individual predictor.

Response Variable

The response variable for our model is **“type of crime,”** defined as a set of 9 crime-type categories consolidated from a subset of the 66 available OFFENSE_CODE_GROUP categories in the raw crime incidents dataset over the years 2016-2019.

The 9 crime-type categories are:

1. Burglary
2. Drugs-substances
3. Fraud
4. Harassment-disturbance
5. Robbery
6. Theft
7. Vandalism-property
8. Violence-aggression
9. Other

Predictors

Listed below are a set of model predictors currently used in our baseline model. While we have additional predictors still under development for future iterations of our model, they were not yet included and are listed separately in Appendix 2 of this document.

1. Day of week

- This is a one-hot-encoded categorical variable for Tue, Wed, Thu, Fri, Sat, and Sun, indicating the day of the week during which the incident occurred.

2. Month of year

- This is a one-hot-encoded categorical variable for Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, and Dec indicating the month of the incident.

3. Night

- This is a binary variable indicating whether the crime occurred between the hours of 8pm and 4am.
- The next iteration of this model will use actual sunset/sunrise times for the date the incident occurred to specify this predictor.

4. Median residential property value

- This is measured by census tract area in which the observation occurred.

5. Median residential value, 3-year CAGR

- This provides a measure of gentrification/development trend activity in the observation's census tract area.

6. Disparity of residential property values (Gini coefficient)

- For this feature "disparity" is measured using the Gini coefficient as a measure of economic inequality in the observation's census tract.

7. Disparity change trend for residential property values (Gini 3-year CAGR)

- This provides a measure of growing or shrinking inequality in the observation's census tract area.

8. Commercial properties mix ratio

- This provides a measure as to how "commercial" the corresponding census tract is, as measured by total assessed commercial property value in the tract divided by the total assessed value for all property in the tract.

9. Commercial properties mix ratio, 3-year CAGR

- Provides a measure of how much more or less commercial the tract is becoming at the time of the observation.

10. Industrial properties mix ratio

- This provides a measure as to how "industrial" the corresponding census tract is, as measured by total assessed industrial property value in the tract divided by the total assessed value for all property in the tract.

11. Industrial properties mix ratio, 3-year CAGR

- Provides a measure of how much more or less industrial the tract is becoming at the time of the observation.

12. Owner-occupied residential property ratio

- What proportion of the residential and mixed-use properties are owner-occupied in the corresponding census tract.
- To a degree this acts as a measure of local ownership as well as a potential indicator of absentee land ownership.

13. Owner-occupied residential property ratio, 3-year CAGR

- Measures trend changes in local ownership for the census tract.

Model Results

While we still have some issues with missingness in several of our features, and instances of fairly strong collinearity with a few pairwise groupings of features to resolve in future iterations of our predictive model (see the collinearity plots and table in Appendix 2 below), our best baseline model, which used Lasso regularization and multinomial classification, resulted in a training accuracy score of 0.2733 and TEST score of 0.2730. And, given the geographic mixing of our response categories and the high bias of these results, we suspect that a linear logistic function is poorly suited for accurately defining our feature space and predicting results. For that reason, we expect to see better accuracy results in future iterations of the model wherein we plan to first use non-parametric methods such as k-Nearest Neighbors and Decision Tree ensemble classification methods, and then later Artificial Neural Networks trained on our feature set, as well as the predictors still under development as outlined in Appendix 1 below.

Even if a logistic function does not provide sufficient expressiveness for our classification problem, it does provide the benefit of interpretable results, from which we can begin to develop a better understanding of the relationships between specific predictors and our response classes. For an overview of these estimated coefficients (as well as an indication of which predictors are found not important via Lasso coefficient shrinkage to zero for certain response classes), please see the figures provided in Appendix 3 below.

Appendices

Below are several additional items referenced in our write-up above, but were moved to the end of the report to allow better flow while reading the document.

Appendix 1: Predictors still under development

The following several predictors were not yet ready for inclusion in our baseline model, but are still being sorted out and will likely be incorporated into our next versions of the model:

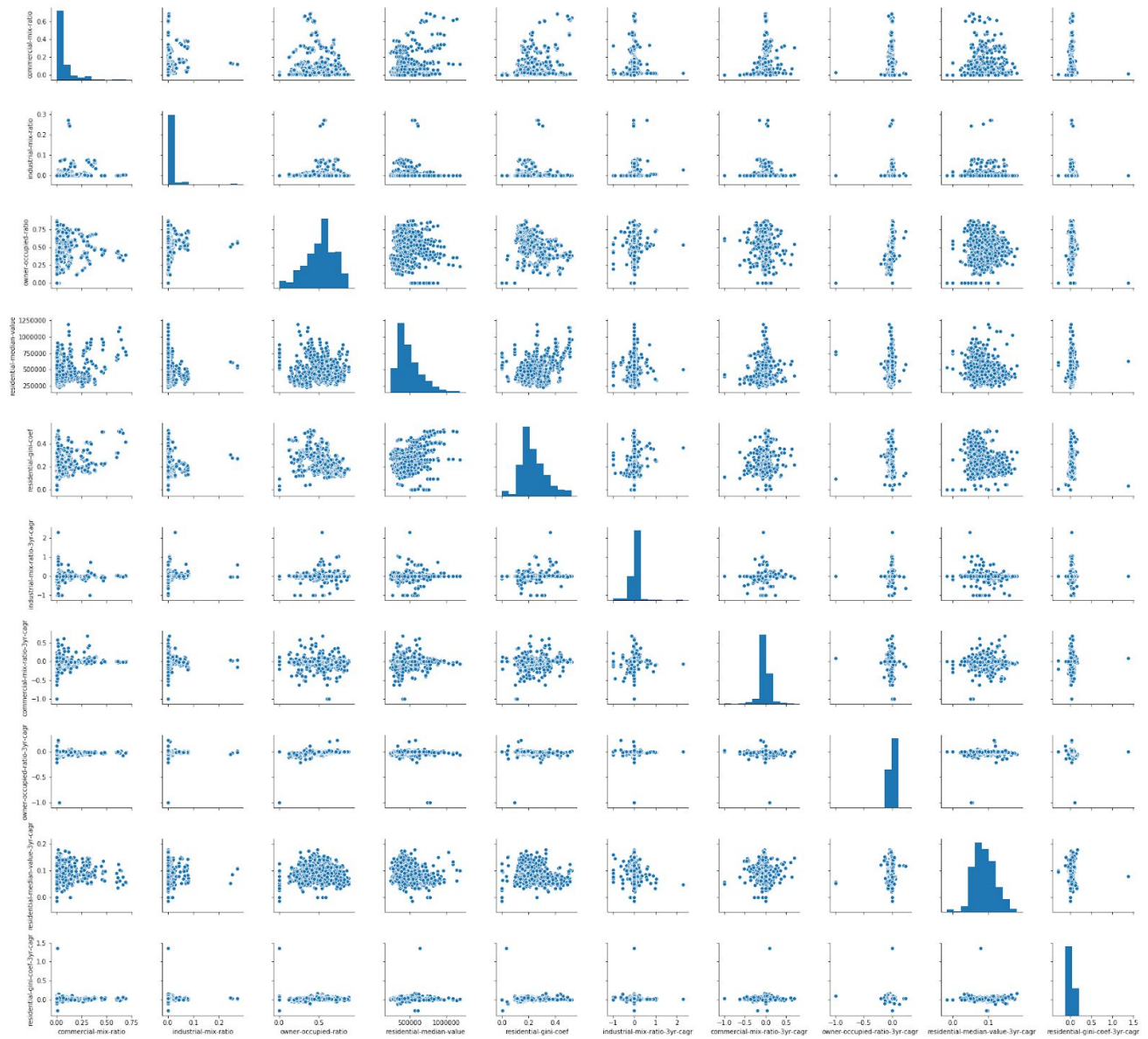
- ***Streetlight lighting density at location of crime***
 - This feature is still being developed and further analysis is needed to determine whether density provides a more meaningful metric when measured by tract or by nearest street segment.
- ***Proximity to university / college campus***
 - Indicates whether the crime occur on or within close proximity to a university or college campus (categorical predictor).
- ***Proximity to high school campus***
 - Indicates whether the crime occur on or within close proximity to a high school campus (categorical predictor).
- ***Liquor license density by tract of observation***
 - We are still considering the inclusion of a census tract liquor license density predictor, however, early EDA indicates that the licensing dataset may not include enough valid data for engineering a reliable feature
- ***Property violations density by tract of observation***
 - We are still considering the inclusion of a property violations density predictor.
- ***Census demographics by neighborhood of observation***
 - We still have some consideration to give as to what demographic features we wish to include. Options include features related to age, poverty, income, and education.

Appendix 2: Overview of predictor collinearity

The most strongly correlated predictors (correlation > 0.20) in our baseline model predictor training set and their corresponding correlation values are:

Predictor 1	Predictor 2	Pairwise Correlation
residential-median-value	residential-gini-coef	0.68
residential-median-value	commercial-mix-ratio	0.58
commercial-mix-ratio	residential-gini-coef	0.57
owner-occupied-ratio	residential-gini-coef	0.42
owner-occupied-ratio	owner-occupied-ratio-3yr-cagr	0.41
residential-gini-coef	residential-median-value-3yr-cagr	0.34
owner-occupied-ratio	commercial-mix-ratio	0.31
residential-median-value	residential-median-value-3yr-cagr	0.28
residential-median-value	owner-occupied-ratio	0.27
owner-occupied-ratio-3yr-cagr	residential-gini-coef	0.26
residential-median-value-3yr-cagr	commercial-mix-ratio	0.24

Pairwise relationships of all numeric (non-binary) predictors in the crime-type training set



Appendix 3: Estimated LASSO regularized multinomial logistic regression model coefficients by response class

