Classifying Parking Violations utilizing Spatiotemporal Data

Harrison Cho
https://github.com/hcho1111

Introduction

To reduce traffic in large metropolitan areas, a road's accessibility, mobility, and openness are top priorities amongst city officials. The pitfalls of lax traffic enforcement extend beyond motorists; long-term negative externalities can affect the flow of commercial goods, air quality, and even pedestrian wellbeing. One aspect of a road's accessibility stems from its parking capacity. As such, complex regulatory schemes have arisen to minimize the externalities of poor driving behavior and general traffic-related inefficiencies. Although dealing with traffic violations in large cities is an issue of significant scale, addressing this problem carries significant monetary and logistical benefits. As such, I propose a data-centric approach to classify traffic violations utilizing vehicle-based, temporal, and location-based features.

I am working with parking violation data from the year 2020 collected by New York City's Department of Finance. The target variable is the parking violation code of a given vehicle[1]. For simplicity's sake, the roughly 100 violation codes are grouped into 17 different categories based on violation similarities[2]. I am working with a subset of my full data; in future analyses, I seek to incorporate multiple years of parking violations. The current dataset is roughly 2.2 GB; there are 12,495,734 observations and 16 features to preprocess. These features range between vehicle, location, and temporal characteristics of individual violations.

Previous analyses have utilized this parking violation data to predict specific aspects of on-street parking. One study by Gao et al. utilized location-based data to determine violation patterns and measure the legality of on-street parking[3]. The study utilized six machine learning classifiers to predict parking legality. The most successful classifier, i.e. the lowest total root-mean squared error, employed random forest classification methods. Overall, their results generalized that commercial, healthcare, and food-service locations were positively associated with increased parking violations. An additional study by Li et al. utilized the same data to predict the availability of street parking within metropolitan areas[4]. Again, random forest models best approximated the availability and timing of parking. The use of spatiotemporal features from the parking violations dataset was coupled with human mobility data and point of interest data. This was done to suggest policy solutions for parking regulators and city managers.

Exploratory Data Analysis

A note on data cleaning procedures is merited before discussing preliminary insights from the data. To conserve space and ease computational time, a number of columns were dropped before preprocessing began. In the future, Brown's distributed computing infrastructure will be used to reincorporate these variables. First, columns whose data had 90%-100% missing values were dropped. Second, because I lacked the effective tools to process this data, data pertaining to specific locations was dropped. This data contained street addresses, cross streets, and street codes. In the future, this data will

[1] Data dictionary listed in the references
[2] Violation codes provided by the NYC department of finance. Link to descriptions and fine amounts provided in references
[3] Gao et al.. "Spatiotemporal Legality of On-Street Parking". 299-312
[4] Li et al.. "Understanding the Spatiotemporal Availability of Street Parking". Sigspatial'19. doi.org/10.1145/3347146.3359366

be run through an API to return latitude and longitude coordinates. Third, due to size constraints, extraneous information, predominantly data of type string, was dropped from the dataset.

A significant part of this analysis stems from classifying parking violations. Utilizing the violation codes and a dictionary provided by NYC's Department of Finance, 100 unique violations were aggregated into 17 classifications. A dictionary for these classifications is in the references section[5]. In addition, some variables were either aggregated or split. For example, I combined prohibited parking time-ranges, originally two columns, into a single column. I split the time a violation was issued, originally one column, into separate month, year, and day columns.

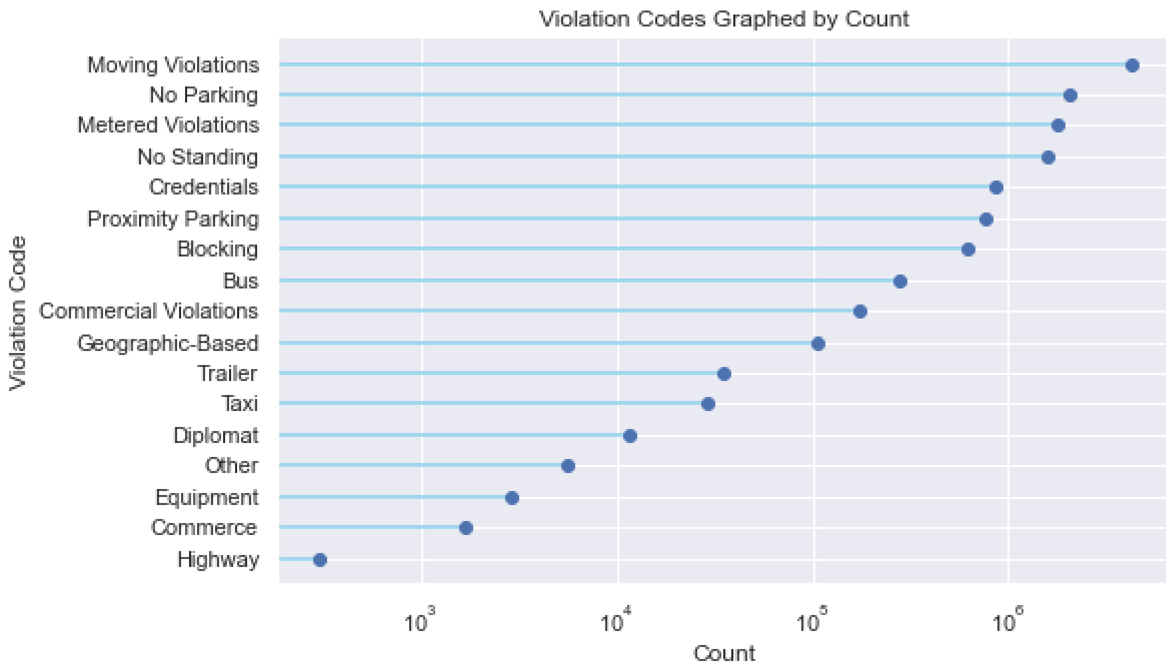

Figure One: The target variable, its groups, and general imbalances between these groups is illustrated. The counts fluctuate significantly given the log scale. Compared to the largest group, the smallest category of violation counts is smaller by several orders of magnitude. In the future, data is normalized when possible to account for this imbalance.

[5] See Appendix Figure A

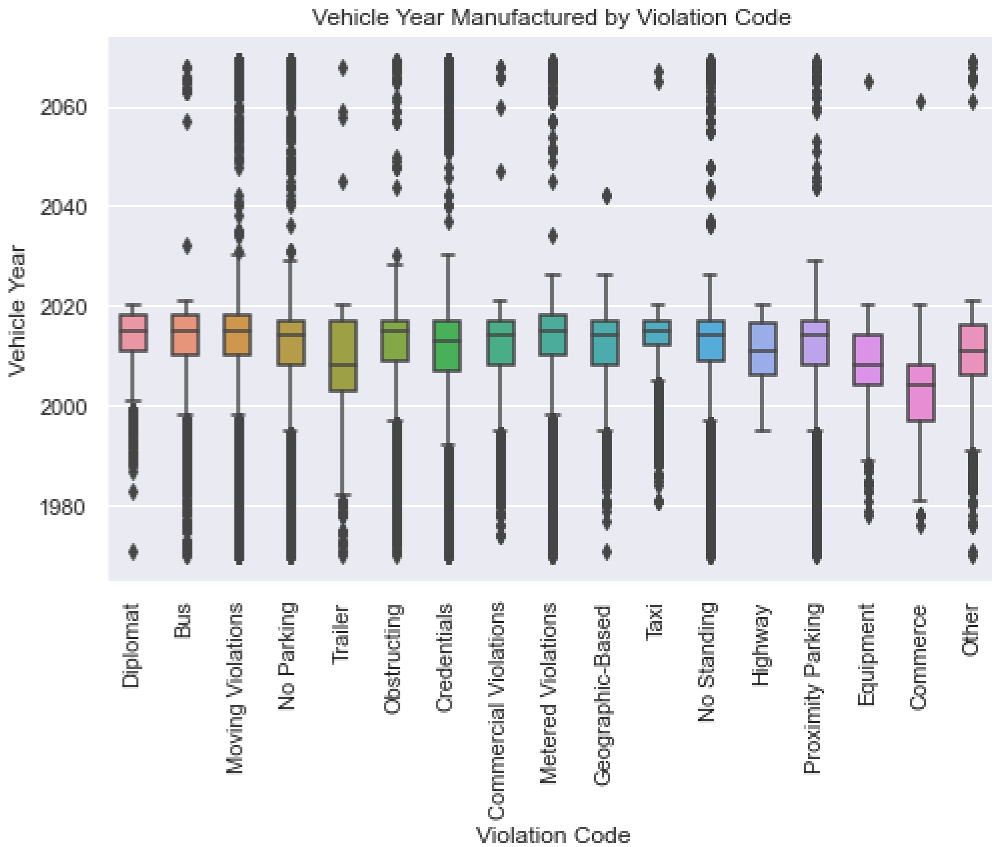

Figure Two: This figure maps the distribution of a vehicle's year to its corresponding violation code. There appears to be many outliers in each distribution, skewing the true group mean. Despite this, it appears that the mean ages of vehicles appear to be similar across violation codes. Significant mean disparities arise for trailer, highway, equipment, and commerce groups. Further investigating whether older cars are more prone to ticketing could be beneficial in future analyses.

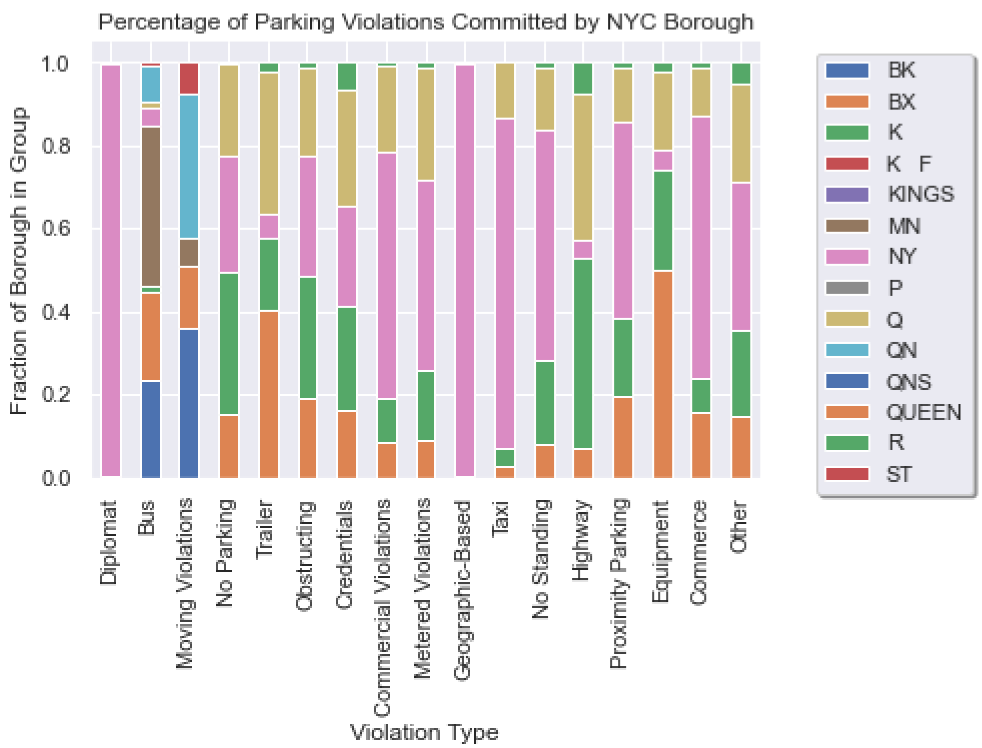

Figure Three: The stacked bar plot illustrates the percentage each NYC borough contributes to a total violation group. Interestingly, both diplomat and geographic-based violations occur exclusively in Manhattan. Overall, it appears that a significant portion of total violations occur in Manhattan.

Not all boroughs appear to be represented in the graph. In addition, borough IDs may not be coded in a standard way. It is also possible that this ID represents different subdivisions of the larger borough. In future analyses, these IDs will be recoded to more accurately reflect borough contributions.

Figure Four: This is a simple time series illustrating the parking violation count by month. Worth noting is the volatility in violation counts between March and June. In future analyses, it may be utile to exclude the year 2020 when running models. Pandemic-era data may not accurately represent parking violation behavior in years where a pandemic is not actively occurring.

Data Preprocessing

Figure One illustrates the imbalances which exist by violation code within the data. In order to account for the discrepancy of counts by violation code, the data is split through the StratifiedShuffleSplit method. It is important to stratify the data such that roughly equal proportions of each violation code appear in each data split. By adjusting for imbalances across the train, testing, and validation sets, we can better predict the class of a parking violation. In the future, when computational power is increased, I plan to incorporate multiple splits in my data for increased generalizability.

I assume that individual parking violations are IID. Intuitively, parking violations are independent of one another, and do not affect the inclusion of other violations. Because individual vehicles are represented, the data lacks group structure. Worth noting are the data's temporal elements. Despite these elements, the data is not a time series because observations are uncorrelated with previous observations.

I have included comments throughout my code that indicate my thought process, and I have utilized random seeds to improve reproducibility. Overall, data clearing and the creation of auxiliary variables is well documented. A word of warning for reproducibility, larger machines are required to run this dataset. Despite removing extraneous variables from my analysis, I was unable to fully preprocess the dataset due to memory allocation issues. In total, my data frame contained 16 preprocessed categorical[6] and numeric[7] features. A list of these features is included in the appendix to illustrate preprocessing decision making. Overall, all categorical features in the dataset do not have a natural order associated with their classification. These categorical features are preprocessed with the OneHotEncoder. Per results from the EDA, all numerical features are unevenly distributed. To address these imbalances, the StandardScaler encoder is employed to normalize counts.

[6] See Appendix Figure B
[7] See Appendix Figure C

Works Cited

NYC Department of Finance. “DOF Parking Codes.” NYC Open Data. Accessed October 10, 2021. **https://data.cityofnewyork.us/widgets/ncbg-6agr.**

NYC Department of Finance. “Violation Codes, Fines, Rules & Regulations.” NYC Department of Finance. Accessed October 10, 2021. https://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page.

Mingxiao Li et al.. “A Data-Driven Approach to Understanding and Predicting the Spatiotemporal Availability of Street Parking.” SIGSPATIAL '19: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2019. doi.org/10.1145/3347146.3359366.

Song Gao et al.. “Predicting the Spatiotemporal Legality of On-Street Parking Using Open Data and Machine Learning.” Annals of GIS. 25:4. 299-312

Appendix

A. Violation Group Code

Original Coding	Renamed	Short Description
100	Diplomat	Parking violations committed by diplomats
200	Bus	Parking violations committed by busses
300	Moving Violation	Violations where person was moving
400	No Parking (Gen)	General No Parking signs present
500	Trailer	Commercial trailers or hauling involved
600	Blocking	Motorist obstructed path
700	Credentials	Improper credentials, plates, or tags
800	Commercial Violations	Violations committed by commercial vehicles2
900	Metered Violations	Parking meters involved in violation
1000	Geographic-Based	Specific geographies with specific rules on parking
1100	Taxi	Violations caused by taxis or other ride-services
1200	No Standing	General no standing violation
1300	Highway	Violations committed on a highway
1400	Proximity Parking	Vehicle proximity violates regulations
1500	Equipment	Improper equipment for vehicle
1600	Commerce	Business/Commerce related violations
1700	Other	Miscellaneous

B. Categorical Features (Involved in Preprocessing Only)

Feature Name	Encoder	Short Description
Registration State	OneHot	Vehicle's home state denoted by license plate
Plate Type	OneHot	Plate class for vehicle
Vehicle Body Type	OneHot	Vehicle body style
Vehicle Make	OneHot	Vehicle manufacturer
Issuing Agency	OneHot	NYC Administrative agency involved in administering violation
Violation Precinct	OneHot	Police precinct where violation occurred
Issuer Precinct	OneHot	Administrative precinct where violation occurred
Violation County	OneHot	NYC borough where violation occurred
Law Section	OneHot	Law subsection responsible for justifying violation
Vehicle Color	OneHot	Vehicle's color
Range	OneHot	Parking zone range of banned times

C. Numerical Features (Involved in Preprocessing Only)

Feature Name	Encoder	Short Description
Vehicle Year	StandardScaler	Vehicle's manufacture year
Distance from Curb	StandardScaler	Vehicle distance from curb
Year	StandardScaler	Year in which parking violation occurred
Month	StandardScaler	Month in which parking violation occurred
Day	StandardScaler	Day in which parking violation occurred