A Closer Look into NY Parking Violations

Cody Hegwer
Applied CS
CU Boulder
Boulder, CO
cohe7798@colorado.edu

Jeff LaPrade
Applied CS
CU Boulder
Boulder, CO
jela2733@colorado.edu

Adam Grabowski
Applied CS
CU Boulder
Boulder, CO
adgr5757@colorado.edu

PROBLEM STATEMENT

New York is notorious for being a difficult place to find parking in. There are millions of people living in New York City, and many streets are narrow and oftentimes one way. By gathering street parking violations from the 1970s until 2014, the goal is to find trends between common violations in order to avoid them. By using the Vehicle Body Type, Vehicle Color, Violation Time, Street, Intersecting Street, and Plate type attributes specifically, our data will guide drivers in New York to be more aware of which types of cars attract attention when marked for parking violations along with certain streets that are surveyed more often than others.

LITERATURE SURVEY

In 2014, I Quant NY published a study analyzing fire hydrants responsible for large amounts of parking violations within New York City using NYC Open Data. The study found that the most ticketed fire hydrant areas had a bike or protected lane that seemed to leave the hydrant open, but oftentimes violations were left to interpretation. Violations from a single hydrant led to $25,000 in fines through one year.

Another study found to be interesting to our project was a study published by Angela Ju on LinkedIn titled “Toronto Parking Ticket Data Analysis.” While this study obviously does not analyze parking in New York, its conclusions helped lead to the motivation in our work. The top 5 violation types in Toronto led to 50% of the revenue from all parking violations. This points to confusion by drivers in how certain streets and parking zones lead to many repeat fines. On top of that, it may point to certain city workers who frequent confusing parking zones knowing that people with mistakenly park there without thinking twice.

By analyzing these two studies, we decided parking violations may not always originate from malintent. Oftentimes using a parking space or not that seems open is a split second decision. So, by finding and marking car types that are often flagged, along with streets and street intersections we hope to help parkers in New York to make better decisions and know the risks of parking in certain areas.

PROPOSED WORK

To begin cleaning the data, there are many attributes which are encoded by a number that need to be translated to usable information. Just to list a few, 'Violation Code' is a 2-digit number referring to why the ticket was issued. There are 3 'Street Code' attributes which are 5-digit numbers, we believe to refer to the location of the violation. There are 'Violation Location' and 'Violation Precinct' which are 3-digit numbers that, from a brief look, tend to be the same.

Another problem is missing values for given attributes. The last 5 columns describe specific violations, such as 'Hydrant Violation' and 'Double Parking Violation'. If we wish to find any interesting information relating to these violations we will have to remove all null rows as the occurrence of said violations are rare, making classification for these pointless. Instead, we would like to classify *given* that the recipient illegally parked in front of a fire hydrant, or *given* that the recipient double parked.

Once the data has been cleaned, we plan to create various predictive models. We will experiment using a subset of our dataset as training data with different attributes as predictors and different classification methods to make the best model. At this stage, it is unclear which attributes will make the best predictors, and examining the accuracy of different models may actually be an answer to some of our initial questions.

Once we believe we believe we have the best model, we may fabricate new data or use another subset of our original dataset to test its predictions. Again, exactly how what attributes will be used in building the model is unclear at this stage: our observations and their relevance to our different inquiries will determine this.

Our inquiries will be very different than of the related work we have identified. We may use classification or predictive modelling in a similar fashion, but ultimately we will be looking answer different questions.

DATA SET

https://www.kaggle.com/new-york-city/ny-parking-violations-issued

Attributes:

Summons Number: Ordinal Integer

Plate ID: Ordinal String

Registration State: Ordinal String

Plate Type: Ordinal String

Issue Date: Ordinal Integer

Violation Code: Ordinal Integer

Vehicle Body Type: Ordinal String

Vehicle Make: Ordinal String

Issuing Agency: Ordinal String

Street Code1: Ordinal Integer

Street Code2: Ordinal Integer

Street Code3: Ordinal Integer

Vehicle Expiration Date: Ordinal Integer

Violation Location: Ordinal String

Violation Precinct: Ordinal Integer

Issuer Precinct: Ordinal Integer

Issuer Code: Ordinal Integer

Issuer Command: Ordinal String

Issuer Squad: Ordinal String

Violation Time: Ordinal String

Time First Observed: Ordinal String

Violation County: Ordinal String

Violation In Front Of Or Opposite: Ordinal String

Number: Ordinal String

House Number: Ordinal String

Street: Ordinal String

Street Name: Ordinal String

Intersecting Street: Ordinal String

Date First Observed: Ordinal Integer

Law Section: Ordinal Integer

Sub Division: Ordinal String

Days Parking In Effect: Ordinal String

From Hours In Effect: Ordinal String

To Hours In Effect: Ordinal String

Vehicle Color: Ordinal String

Unregistered Vehicle?: Boolean String

Vehicle Year: Ordinal Integer

Feet From Curb: Quantitative Integer

EVALUATION METHODS

Predominantly, we will be analyzing the accuracy and precision of the predictive model to answer our questions. Due to the nature of these inquiries, the main events we wish to predict are future events that, for the sake of this study, we would not have access to. This means that the quality of our model is the main concern of our evaluation. All of the knowledge we wish to gain is based on the prediction, and so its effectiveness very well may answer our questions as well as being our reference point. The precision of our classification will be more readily evaluated since we may predominantly be testing on data already in our dataset as opposed to new, more recent data.

TOOLS

There are two main platforms we wish to explore our data with. Tableau for visualization and Python for raw data calculations, statistics, and classification. Specifically, we'd like to utilize Numpy, Pandas, and Matplotlib libraries in Python.

1 Tableau

Tableau is a tool for rapid creation for data visualization, allowing the user to input any data, typically a .csv, and generates graphs based on user selected attributes.

A major benefit to using Tableau is its ability to quickly and easily map locations globally, identify borders, and fill in these areas with desired attributes. For example, given US counties voting results for the 2016 presidential elections, Tableau automatically identifies all the counties locations on a map and, with user instruction, fills in the counties with red and blue based on the counties voting preferences.

This seems interesting to us, as we would like to see if we could discover some trends if the data was mapped to locations, ideally street addresses, in New York City. There are two glaring limitations to this though. One, is if we find that the street addresses are too general and only reduce to the actual street name or we find that majority of tickets issued provided no information for this attribute. Second, we are unfamiliar with if Tableau's location mapping can actually handle street names.

2 Python

The programming language we decided to use for our actual data mining is Python. It is the language we are most familiar with in our college career. It is simple and a good tool for data manipulation. Python also has many well flushed out libraries intended for use by data scientists. The following will be used extensively.

2.1 Numpy/Pandas

Numpy and Pandas are the main libraries being used for the project. Both contain many statistical measures and tools. Additionally, Pandas contains data structures to convert .csv files into easily manipulatable tables.

2.2 **Matplotlib**

The Matplotlib library contains many options for graph and chart creation. It pairs nicely with Numpy and Pandas as the data we are concerned about are converted to usable arrays.

MILESTONES COMPLETED

There are 5 remaining project turn-ins before the end of the semester. 2 of which allude to the completion of the project. This section is dedicated to setting a timeline for ourselves to abide by as well as giving an outline of the work for following weeks.

1 **Project Progress Report**

The next major assignment is a Project Progress Report. It should be an update to this document, specifically an update on the milestones we've completed. The following are goals we would like to complete for this assignment:

1.1 **Confirmation Meeting(s)**

Brief meetings to confirm details about the data set, which attributes we wish to focus on, what hypotheses we hope to prove, etc. Should be regular to keep other milestones on track.

1.2 **Data Cleaning**

After analyzing the full dataset, it became clear that before the years 2013 and 2014, there were many holes in data collection. With that in mind, our team decided remove those prior years, determining that 2013-2014 still contained millions of points of data that were relevant and complete enough to sustain our study.

To clean the data, once scope was reduced to 2013-2014, we worked to remove erroneous data points with misspellings or holes of the attributes selected for the study. One attribute specifically, Violation Time, needed to be manipulated from the AM/PM format into a 24 hour standard measurement to allow easier comparisons to be made. Additionally, spell checks were implemented for various attributes to correct commonly misspelled words such as the color "black" being spelled as "balck." There were many color abbreviations than expected. See short example below:

'WINE' 'WOOD' 'YELLL' 'YWG' 'TRQ' 'LEXUS' 'BK/WH' 'GY/W' 'GY/WH' 'GY/WT' 'BLWHI' 'PETER' 'W UIZ' 'W/RED' 'WT/GN' 'GML' 'RD B' 'BLK/O' 'BRU'

We created buckets for major colors like red, green, black, etc. Depending on what letters the abbreviations started with we located them in an appropriate bucket.

Similarly for which state the vehicle was registered in, we used only tuples containing valid state abbreviations.

The following two distributions give a view of the number of entries based on the time of violation and color of vehicle.

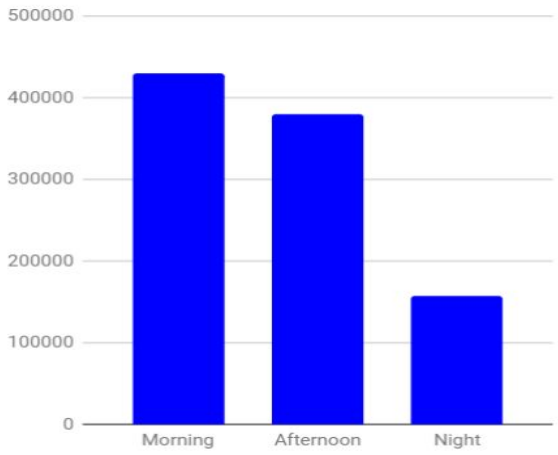

Fig 1: Distribution of data based on time of violation

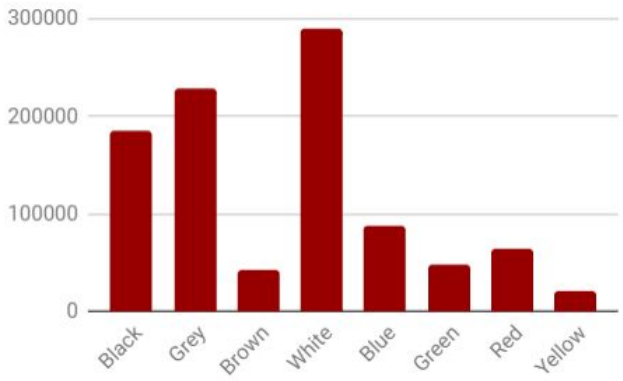

Fig 2: Distribution of data based on color of vehicle

1.3 **Model Building and Analysis**

With the data cleaned, we begin to build a classification model. Although we plan to include additional models later in the semester, we initially started with a Bayesian classifier. Although not one of our originally stated problem statements, we decided to build a model to predict color of the vehicle, just out of our own curiosity. We built the model to predict color of vehicle based on: violation code, time of violation issue, and state of violation issue as an initial test run. Now that we have our basic Bayesian classifier built, it will be simple to add or remove additional attributes to build different Bayesian models as we see fit.

Our next step is expand the Bayesian model with additional attributes as well as experimenting with different combinations to attempt to identify a strong predictor. Afterwards, we plan to build an Apriori algorithm which we believe could provide some interesting insight.

MILESTONES TO-DO

1.3.2 **Model Building and Analysis-Part 2**

The next step is to continue our Bayesian analysis of the data and additionally move into other classification methods as well. Our first plan is experiment with differing combinations and varying amount of attributes to attempt and identify the strongest Bayesian classifier.

Since we have our initial framework for constructing Bayesian models put together, constructing additional models from this point will be very simple and much less time consuming than up to this point. Therefore, over the next few weeks we plan to perform the bulk of our analysis.

An Apriori algorithm is the next type of classification we plan to explore, after concluding our Bayesian analysis. Although Apriori is typically used for market-basket research, we believe that it may provide us some interesting insight given our problem space and the questions we wish to answer. Not only is our data “transaction-based”, which Apriori is well suited for, but additionally, many of our problem statements wish to consider frequent combinations of attributes such as: which streets and vehicle models are frequently put together in violation transactions.

Even if the Apriori analysis does not in itself answer all our questions, we believe that, if nothing else, it will strongly contribute to the other knowledge we will have gained from our other models.

1.3.3 **Model Building and Analysis - Visualization**

Once we have created the majority of our models, we plan to create visualizations of our models' prediction results. Specifically, the results of our models attempting to classify testing data. Although our intent is that much of our knowledge will be readily gained from the numerical results alone, it is possible that the visualizations will cause us to notice a trend that we would miss otherwise and so realize that it is imperative to also explore the data in this way.

Since we are dealing with classification, we primarily plan to view the data through scatter plots, possibly also with our model superimposed, to be able to try and clearly see how our models are classifying and separating the data. Scatter plots are commonly used when analyzing predictors for that reason: it is simple for both the creators and outside parties to at-a-glance identify how data points are categorized based on color and proximity in the graph.

2 Final Project Report

When our project is finally completed, we will write up our final results and finding. The following milestones represent outstanding tasks that are not scheduled to be completed until after our analysis is actually done and after our Final Project Report:

2.1 Cleanup/Commenting of Code

Simply label and organize the code to make it easier for third-parties to read and understand our methodologies.

2.2 Prepare Presentation

Our final presentation will most likely be created through Powerpoint and will include not only our problem statements, how we approached them, and examples of our code, but also any relevant visualizations that may help viewers to understand our findings more easily or ones that were important to our own understanding or knowledge gain.

Results So Far

At this stage, we have primarily spent our time constructing our initial code and model creator. As such, we have not done as much analysis as of yet since we have primarily been focused on identifying how we wished to create and organize our code. It is important to us that we use many different models since we are not entirely certain which combinations of attributes we expect to be the strongest predictors. As a result, we wanted to take the time to structure our code so that it not only would allow us to simply create additional models in the future, but also give us the flexibility to swap between attributes and create models with many different configurations.

Despite the bulk of our analysis coming in the next steps, we have made some initial observations. So far we have implemented a naive Bayesian classifier, given 3 attributes, we can predict what the color of the vehicle. The following results were found for 3 different tests.

For a vehicle with plates from New York, with violation code 20; 'General No Parking', and violation issued at **night** time:

1. White : (prob = 3.66e-05)
2. Grey : (prob = 2.92e-05)
3. Black : (prob = 1.12e-05)

For a vehicle with plates from New York, with **violation code 40**; 'Stopping, standing or parking closer than 15 feet of a fire hydrant', and violation issued in the morning:

1. White : (prob = 1.10e-04)
2. Grey : (prob = 5.49e-05)
3. Black : (prob = 2.10e-05)

For a vehicle with plates from **New Jersey**, with violation code 20; ‘General No Parking’, and violation issued at night time:

1. White : (prob = 1.93e-05)
2. Grey : (prob = 7.61e-06)
3. Black : (prob = 3.45e-06)

From our early observations, it appears that color will most likely not have any correlation to the other three attributes as the outcomes have been the same for all tests and seems to only revert back to the distributions in figure 2.

We may attempt to sample the data so that we create an even distribution across all colors, and then try to classify in hopes to prove color is either correlated / uncorrelated.

Additionally, we plan to further develop the classifier to be able to classify a vehicle on the other 3 attributes.