Montgomery County Traffic Violations Data:

An Exploratory and Predictive Study

Patrick Aquino, Adam Imran, Thomas P. Malejko, and Douglas Post

Georgetown University

Montgomery County Traffic Violations Data: An Exploratory and Predictive Study

Abstract—To Be Written

**Background**

To be Written

**The Data**

**The Raw Data**

The principal data used in this research comes from the ‘Traffic Violations’ dataset, which is part of Montgomery County Maryland’s Digital Government Strategy. This dataset contains information about every electronic traffic violation issued in Montgomery County, Maryland from January 1, 2012 to March 4, 2020 (the date of collection for this study). The dataset, in its original form, contains 1.66 million records spanning 43 features—or more than 71 million individual data points. The features include general information about the stop itself (i.e. date of the traffic stop, a description of the violation, and location), data about the vehicle involved (i.e. model year, make, and color), and demographics about the driver (i.e. race, gender, state of residence). Features that could be used to identify the specific vehicle, its operator and/or owner, or the ticketing officer were removed by the county prior to the dataset’s publication (*Traffic Violations,* March 4, 2020).

Supplemental data used in this research comes from Visual Crossing Corporation’s Weather Forecast and Historical Weather Data API. This dataset contains information about hourly weather conditions for Gaithersburg, Maryland from January 1, 2017 through December 31, 2019 and contains information about the temperature, humidity, precipitation, wind speed, etc. for the specified location (*Weather Forecast*, March 12, 2020). In total the retrieved dataset from this source has 26,258 records spanning 16 features.

A detailed description of the data (and its features) as applied in this study can be found in Appendix A (Raw Data).

**Data Wrangling and Munging**

As stated above, the Montgomery County Traffic Violations dataset contained more than 1.66 million records spanning 43 features; however, that expansive dataset was too large to process using the techniques learned in Georgetown’s Analytics 512 Course. As a result, the dataset was abridged to a three-year period—from January 1, 2017 to December 31, 2019. The resulting dataset that was still large, but small enough to apply the techniques learned in this course using local computing. Once reduced in size, weather data from the Weather Forecast and Historical Weather Data API was integrated into the traffic violations dataset to provide a more comprehensive group of features that would allow for a detailed assessment about the effects of environmental factors on community policing and traffic incidents. The integration of weather data resulted in one change to the original—traffic violations—dataset, which was the loss of minute-specificity in the *Time.of.Stop* feature.

The data wrangling process was fairly intensive for this dataset, requiring the removal or modification of several columns due to unclear definitions and large quantities of missing or erroneous values. The table below summarizes the columns removed from the original dataset and the reasons.

|  |  |
| --- | --- |
| Feature Name | Reason for Removal |
| Traffic Violations Dataset | |
| *Date.Of.Stop* | Merged with the *Time.Of.Stop* to form a consolidated *Date.time* feature |
| *Time.Of.Stop* | Merged with the *Date.Of.Stop* to form a consolidated *Date.time* feature |
| *Agency* | Limited values—all records contained ‘MCP’ |
| *Geolocation* | Redundant—contained data about the *Latitude* and *Longitude*, which are individually listed in their own column |
| *Search.Outcome* | Reformatted to indicate whether an arrest occurred due large quantity of missing values. Now called *Arrest.* |
| *Search.Reason.For.Stop* | Large quantity of missing values |
| *Search.Arrest.Reason* | Renamed to *Arrest.Reason* |
| Weather API Dataset | |
| *Location* | Unnecessary. All records contained Gaithersburg, Maryland |
| *Resolved.Address* | Unnecessary for this assessment |

*Table 1. Features removed from the dataset and the reasons*

The original merged dataset contained 9.27% missing or erroneously recorded values, which was too large for any future analysis. After the dropping the columns specified in *Table 1* and conducting additional munging, the dataset cleanliness increased to 92.27% with no values occurring outside their specified range. While this number may still seem large, consider that preponderance of this remaining value are ‘true’ missing values. For example, if a pedestrian is issued a citation for jay-walking there would be no model year recorded since a vehicle was not involved. Similarly, if a search was not conducted pursuant to the traffic stop, there could not be a reason for the search. As a result, the final dataset as applied in this research has an effective cleanliness in excess of 99%.

**Exploratory Data Analysis**

To be Written

**Visualization of the Problem**

To be Written

**Feature Generation**

To maximize the utility of all columns provided in the original dataset, several new features were generated. The sections below, detail any modified features.

***Traffic Stop Location***

The *Location* column of the dataset is a free-text field, in which the citation issuing authority uses common street names to identify the location of any traffic stop (i.e. ‘Germantown Rd @ Crystal Rock Dr’ or ‘4715 Cordell Ave’). While there are some repeat locations, the majority of columns are unique and, as such, cannot be used a categorical feature. In an attempt to maximize the amount of information derived from the data, two new features were generated from this information: *Highway* and *MajorRoad.* These new features support evaluations of a hypothesis that officers operate differently on commonly traveled sections of road, compared to local streets and residential areas. Therefore, these generated features are bools that indicate whether the incident occurred on a highway or major route based on a regular expression search of the *Location* feature. Based on the geography of Montgomery County, I-270 and I-495 were the only roads determined to be a ‘highway.’ Major roads are high-traffic, multiple-lane routes that are commonly known as state-highways or county routes. In Montgomery County, the following roads (and their associated common name such as Old Georgetown Road or Wisconsin Avenue) were determined to be ‘major roads’: I-270, I-495, US-29, MR-27, MR-28, MR-29, MR-97, MR-112, MR-185, MR-187, MR-190, MR-191, MR-193, MR-200, MR-320, MR-355, MR-390, MR-410, MR-500, MR-586, MR-650, and 16th Street.

***Traffic Infractions and Their Specific Charges***

The *Charge* feature provides a specific law in the Maryland Transportation Code, which details the exact section of the state’s code that the driver violated. For example, a charge of ‘21-201(a1)’ corresponds to a driver who failed to obey a traffic control device (like a stop sign or red light). While the information is very specific, there are more than 890 unique entries, which limits the column’s ability to be used as a categorical feature. Since the information in this column can be very valuable in a number of predictive or inferential analyses, we generated a new feature—*ShortCharge—*that only considers the title and subtitle of the Maryland Code in question. From the example above, ‘21-201(a1)’ becomes ’21-1’ in the *ShortCharge* feature, which corresponds to ‘Vehicle Laws – Rules of the Road—Traffic Signs, Signals, and Markings.’ While this may seem overly broad, the resulting feature still has 49 unique levels. Note, during random forest modelling—where the maximum number of levels is 47—this feature was reduced to title only (i.e. ‘21’ or ‘Vehicle Laws – Rules of the Road’).

***Maximizing the Value of Date and Time Information***

To extract the maximum value from the *Date.time* feature, two new columns were generated: Day of Week (*DoW)* and *Hour.* During exploratory data analysis, these new features had significant differences regarding the rate of citation-to-warning and, therefore, will allow for unique hypotheses testing about policing tendencies relatively to the incident time and day of the week. Note, there was no evidence to support a change in policing relative to the time of month (beginning v. end of the month); subsequently, no column was generated to evaluate this feature further.

***Multiple Infractions***

It seems only natural, but individuals violating multiple rules of the road (i.e. speeding, failing to stop at a red light, and possessing an expired license) may be treated differently by police officers than those who only commit a single violation. Therefore, the column *MultiInfr* was generated to identify traffic stops in which the driver violated multiple portions of the Maryland Transportation Code. This bool was generated when the same *SeqID* was present for multiple records.

***Accounting for the Effects of the Maryland, Virginia, and District of Columbia Area***

While the original dataset provided a detailed record about drivers (and their vehicles) to the state-level, the number of factors was simply too great. As a result, *State* and *DL.State* were transformed to indicate whether the driver, or vehicle, was from the Maryland, Virginia, or Washington D.C.-- this new feature consists of a two-level factor comprised of ‘DMV’ or ‘Other.’

**Methodology**

**Statistical Analysis**

To be Written

**Linear Models and Comparison**

To be Written

**Results and Interpretations**

To be Written

**Conclusions**

To be Written

References

*Traffic Violations*. (March 4, 2020). dataMontgomery. Retrieved March 4, 2020, from https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q#

*Weather Forecast & Historical Weather Data*. (March 12, 2020). Visual Crossing Corporation. Retrieved March 12, 2020, https://www.visualcrossing.com/weather-data

Appendix A

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Format** | **Description** |
| SeqID | String | Unique identifier for each traffic stop (multiple rows can have the same SeqID if multiple citations were issued, etc.) |
| Date.time | POSITIX | Date and time of traffic stop (time rounded to the nearest hour) |
| SubAgency | Factor | Court code representing the district of assignment of the officer |
| Description | String | Text description of the specific charge |
| Location | String | Text description of the violation (usually an address, intersection, or highway exit) |
| Latitude | Float | Latitude location of traffic violation |
| Longitude | Float | Longitude location of traffic violation |
| Maximum.Temp… | Float | Maximum temperature on the day of the traffic stop |
| Minimum.Temp… | Float | Minimum temperature on the day of the traffic stop |
| Temperature | Float | Temperature at time of the traffic stop |
| Wind.Chill | Float | Windchill at time of the traffic stop (if applicable) |
| Heat.Index | Float | Heat Index at time of the traffic stop (if applicable) |
| Precipitation | Float | Total precipitation during the hour of the traffic stop |
| Snow.Depth | Float | Snow depth at time of the traffic stop |
| Wind.Speed | Float | Average wind speed at time of the traffic stop |
| Wind.Gust | Float | Maximal wind gust at time of the traffic stop |
| Cloud.Cover | Float | Average cloud cover at time of the traffic stop |
| Relative.Humidity | Float | Average relative humidity at time of the traffic stop |
| Conditions | Factor | Description of weather conditions at the time of the stop |
| Accident | Factor | YES if traffic stop involved an accident |
| Belts | Factor | YES if seat belts were used in accident cases |
| Personal.Injury | Factor | YES if traffic violation involved personal injury |
| Property.Damage | Factor | YES if traffic violation involved property damage |
| Fatal | Factor | YES if traffic violation involved a fatality |
| Commercial.License | Factor | YES if driver holds a commercial driver’s license |
| HAZMAT | Factor | YES if traffic violation involved hazardous material |
| Commercial.Vehicle | Factor | YES if vehicle committing the violation is a commercial vehicle |
| Alcohol | Factor | YES if traffic violation included an alcohol related suspension |
| Work.Zone | Factor | YES if traffic violation was in a work zone |
| Search.Conducted | Factor | YES if a person or property search was conducted |
| Search.Disposition | Factor | Resulting outcome of the search |
| Search.Type | Factor | Type of search conducted (person, property, both, etc.) |
| State | Factor | State issuing the vehicle registration (including Canadian Provinces and US Territories) |
| Search.Reason | Factor | The reason for the search (Probable Cause, Warrant, etc.) |
| VehicleType | Factor | Type of vehicle involved in the traffic stop (automobile, light truck, motorcycle, etc.) |
| Year | Int | Year the vehicle was made |
| Make | String | Manufacturer of the vehicle (Ford, Lexus, Mack Truck, Indian, etc.) |
| Model | String | Model of the vehicle |
| Color | Factor | Color of the vehicle |
| Violation.Type | Factor | Violation type: Warning, Citation, or ESERO (Emergency Safety Equipment Repair Order) |
| Charge | String | Numeric code for the specific charge (legal citation) |
| Article | Factor | Article of state law (TA = Transportation Article, MR = Maryland Rules) |
| Race | Factor | Race of the driver |
| Contributed.To.Acc... | Factor | YES if traffic violation was contributing factor to the accident |
| Gender | Factor | Gender of the drive |
| Driver.City | String | City of the driver's home address |
| Driver.State | Factor | City of the driver's home state |
| DL.State | Factor | State issuing the driver's license |
| Arrest | Factor | Did the traffic stop result in an arrest (TRUE if yes) |
| Arrest.Reason | Factor | Reason for the arrest |
| Asset.Type | Factor | Type of asset used to generate the citation (A=Marked Car, Q=Marked Laser, etc.) |

Appendix B

R Code Script

To be Written