Traffic Violations:

An Exploratory and Predictive Look at Traffic Incidents in Montgomery County, Maryland

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

Georgetown University

Analytics 512: Statistical Learning

Dr. Purna Gamage & Dr. Keegan Hines

30 April 2020

It happens to just about every American over the course of their life; at some point they are stopped by the police for violating a rule (or rules) of the road. However, what happens after that point is largely up to the citing officer—does the driver get issued a citation or are they simply given a warning? Will the officer's decision be based purely on the severity of the driver's violation or do environmental factors impact their decision? At a more malicious level, does the officer's decision depend upon their preconceived notions of justice or their personal biases towards the vehicle's operator? While some of these questions may seem innocuous, understanding how law enforcement officers apply discretion is nontrivial. Due to the amount of trust afforded to them, the decisions made by individual officers can have societal-wide implications. While challenging and time-consuming to understand individual decision-making processes, the Montgomery County (Maryland) Traffic Dataset allows for an aggregated look at how this police-force—collectively—makes decisions. Using a smorgasbord of analytical techniques and machine learning processes, we determined that Montgomery Police Officers generally issue citations based upon the severity of the traffic violation and other incident characteristics (i.e. was there an accident). Environmental factors (such as weather and time of day) may also impact an officer's decision to issue a citation or warning, while driver demographics appear to have little to no impact.

Research Background and Motivation

Law enforcement officers are entrusted with great responsibility and discretion as their duties involve them operating unsupervised, with fleeting public oversight at most. As a result, an officer's effectiveness is largely dependent upon the trust that they, and their fellow officers, have built with the community. When that trust is broken, the effects can have devastating social and economic consequences as the Black Lives Matter Movement demonstrated (Baumgartner et

al., 2017). Therefore, every officer interaction with a private-citizen is an opportunity to build and maintain that trust with the public. Since the most common form of officer-citizen interaction is the traffic stop (*Bureau of Justice Statistics—Traffic Stops*, n.d.), the public availability of this information allows for a unique look at local law enforcement's efforts in this crucial field.

The purpose of any traffic stop is to enforce traffic regulations and public safety during a potentially hazardous activity that resulted in 36,560 fatalities across the United States in 2018 (*Roadway Fatalities*, 2019). During these engagements, officers are entrusted with a great deal of discretion that includes determining the incident's outcome, whether it be a 'no action,' warning, citation, emergency service repair order (ESERO), vehicle or personal property search, or arrest. This range of outcomes carries a myriad of second-order effects, from nothing to financial penalties and criminal convictions, which makes each traffic stop an important and sensitive matter for the vehicle operator. Therefore, controversies about racial profiling, weekly or monthly effects (i.e. where officers may issue more tickets towards the end of the month), and other demographic influences during traffic stops tend to exacerbate any distrust between the public and law enforcement (Liu & Sharma, 2019, p. 1).

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

Through a thorough exploratory analysis, many interesting things can be discovered. As discussed above, the data covers 43 feature variables over a three-year period, from 2017 to 2019. Associated with those variables are categorical and numerical features. The numerical features are typically related to meteorological information, while the categorical variables comprise the majority of the data. Variables like *Arrest*, *Accident*, and *Search Conducted* have values in the form of 'Yes' or 'No.' To explore the categorical variables, the most common value was identified while the minimum, mean, and maximum values were calculated for any numerical feature. The tables below detail these summary statistics. Please note that any variable not listed was a binary feature, which all had a mode of 'No' or "False' respectively.

Feature	Mode Value (Most Violations)
Subagency	Wheaton MD.
Description	Failure to Obey Traffic Sign
Search.Type	Both (Person & Property)
State	MD
VehicleType	Automobile
Make	Toyota
Color	Black
Violation.Type	Warning
Charge	21-801.1 (Speeding)
Race	Black
Gender	Male
Asset.Type	Marked Patrol

Feature	Min	Mean	Max
Temperature	1.20	56.43	97.9
Wind.Chill	-12.4	25.4	49.1
Heat.Index	78.9	88.7	111.4
Precipitation	0	0.002	3
Snow.Depth	0	0.003	7.1
Wind.Speed	0	5.37	32.3
Wind.Gust	0	1.77	71.4
Cloud.Cover	0	39.73	100
Relative.Humidity	13.67	68.27	100
Year	1929	2009	2020

Table 2. Summary statistics for all pertinent numerical and categorical variables

Diving deeper into the data, a few interesting trends become apparent such as the most common reason for arrests being Driving Under the Influence (DUI) of Alcohol. Additionally, there were 15,654 accidents, 157 of which resulted in at least one fatality, on Montgomery County roads between 2017 and 2019. In terms of this study's principal objective, there were 347,361 warnings and 231,025 citations issued by the Montgomery County Police Department during this study's timeframe. This presents a series of questions like: Which variables above are most likely to cause a driver to be issued a citation versus a warning? Is a particular person more likely to get a citation? Do environmental factors (weather and location) contribute to an officer's decision?

The first step in understanding these questions is to analyze the probabilities of certain binary variables and their effect on citation rates. The table below summarizes this result:

Factor	P(Citation | Factor)
Alcohol = 'Yes'	93.96%
Accident = 'Yes'	92.86%
Property.Damage = 'Yes'	89.67%
Search.Conducted = 'Yes'	78.00%
Workzone = "Yes'	66.94%
Fatal = 'Yes'	58.59%
Belts = 'Yes'	56.00%
Highway= 'True'	48.08%

Table 3. Conditional probabilities table for an affirmative binary outcome

Based on the results from this preliminary analysis, it is clear that there are multiple lines of inquiry worth investigating as there are clearly some factors that contribute to a higher percentage of citations than others.

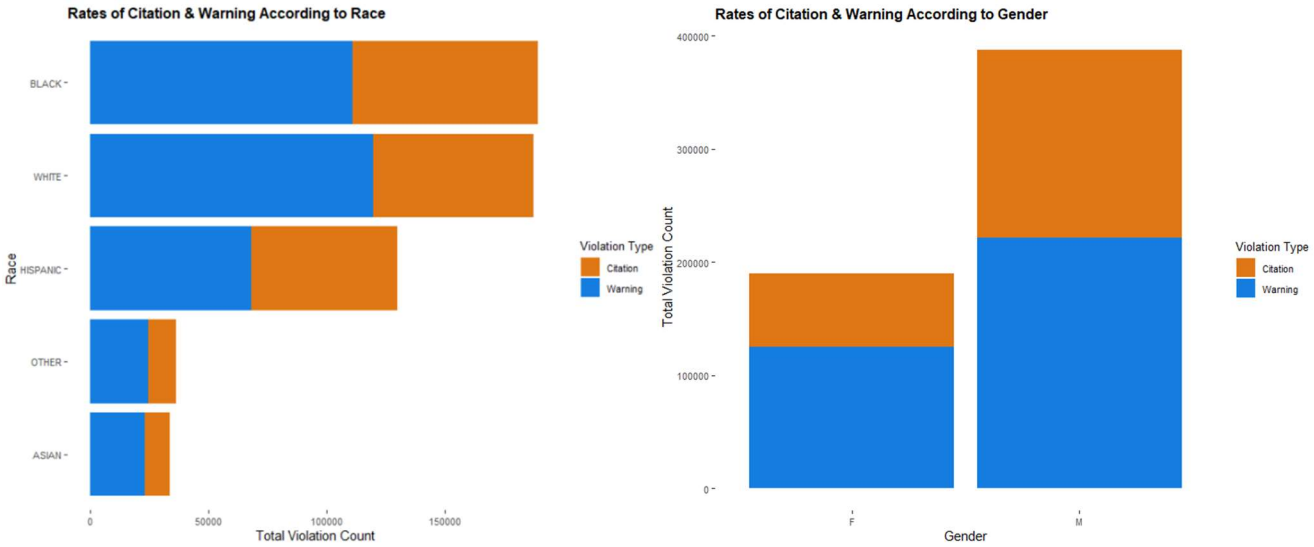

Figure 1. Two charts highlighting potential relationships between driver demographics and rates of citation.

The charts above also show certain factors may have an effect on a traffic stop's outcome, however more analysis is needed before any conclusions can be drawn about these factor's actual effects. For example, it appears that male drivers receive citations at a higher rate than female drivers however this observation could be caused by male drivers committing more serious offenses than their female counterparts. As a result, further analysis is needed before any firm conclusions can be drawn about the effects of various factors on a traffic stop's outcome.

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 factor levels is 47—this feature was reduced to the 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 indicated different citation-warning rates and, therefore, should allow for unique hypothesis testing about policing tendencies relative 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.

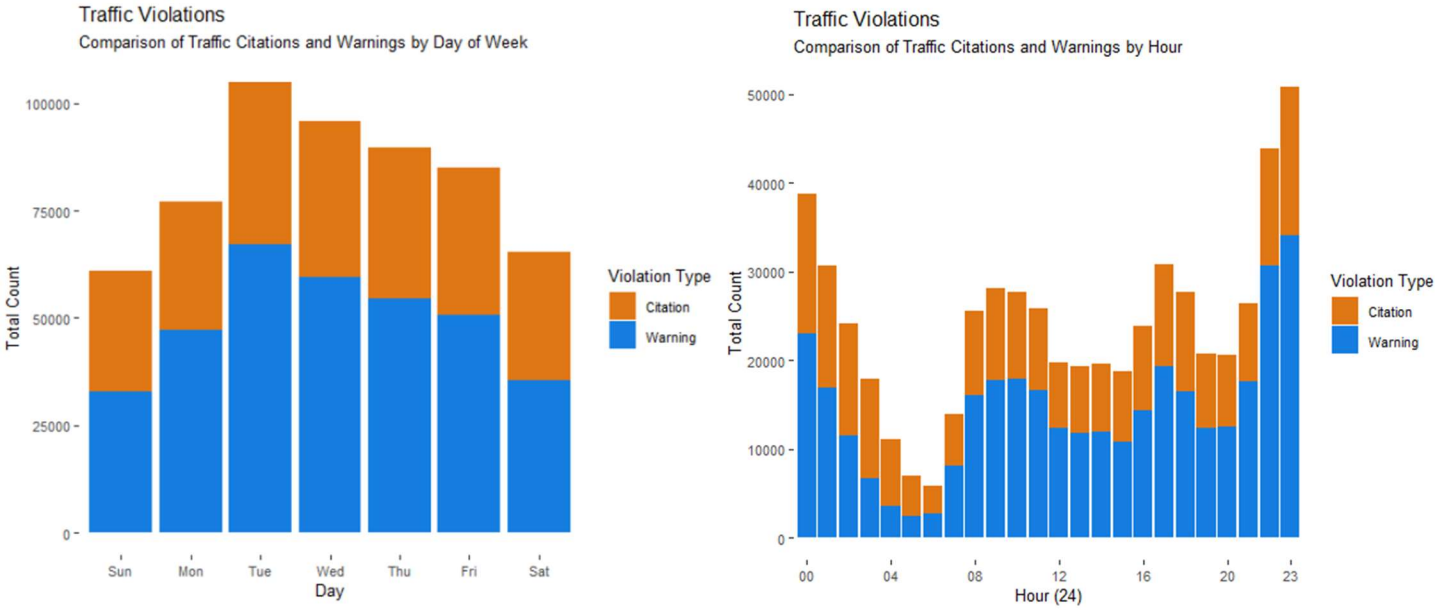

Figure 2. Understanding the effects of the day of week and hour of traffic incident have on citation rates.

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 boolean was generated by finding instances where multiple records have the same *SeqID*.

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 Maryland, Virginia, or Washington D.C.-- this new feature consists of a two-level factor , 'DMV' or 'Other.'

Methodology

Research Objectives

Based upon the exploratory data analysis, background research, and relevant domain knowledge, the following research objectives were identified. These objectives were thoroughly evaluated using various analytic techniques and machine learning algorithms as to provide maximum prediction capability while, simultaneously, understanding the impact of each feature on the traffic stop's outcome.

1. Evaluate the available data to predict a traffic stop's outcome (whether the driver receives a citation or warning). Similarly, understand the factors and variables that significantly influence an officer's decision.
2. Evaluate, and understand the factors that influence, the likelihood of a vehicle occupant incurring an injury due to an accident, given relevant factors.

Machine Learning Methodology

The aforementioned objectives were principally evaluated using various machine learning methods and techniques. This approach was somewhat complicated by the large number of categorical variables in the dataset—more than half—with many having upwards of 50 levels.

As a result, Principal Component Analysis, or similar techniques, could not be used to hone the feature space to a more ideal size. Therefore, a Random Forest ensemble was constructed for the entire dataset to better understand it, particularly the relative importance of each variable in predicting the *Violation.Type* ('citation' or 'warning'). Any features with extremely low variation or prediction capability would be dropped from any future analysis in order to save time and computational effort. The reduced dataset would then be tried against a series of machine learning techniques to determine the model with the highest test prediction score. Each model was tuned appropriately to ensure an optimal combination of hyperparameters before rendering a decision about a particular model's utility. The techniques evaluated include: General Boosted Models, Random Forest, Lasso Logistic Regression, Best Subset Selection, and Hierarchical Clustering. The large size of the dataset incurred additional challenges, which necessitated the use of the 'Caret' package, which enabled many of these algorithms run-in-parallel across multiple cores on a single machine (Kuhn & et al, 2020).

Statistical Analysis

Additionally, traditional statistical methods were applied, as required, to understand and certify the significance of certain trends that emerged in the data.

Results and Interpretations

Random Forest (Feature Selection)

This initial random forest ensemble provided a mechanism for feature selection by identifying variables with extremely low prediction capability. Consequently, the model was built across the entire dataset, using a modest set of hyperparameters (*mtry* = 4 and *ntree* = 100). Prior to constructing the model, unique features (i.e. *SeqID* and *Description),* geolocation features (i.e. *Latitude),* low-variation fields (i.e. *Arrest.Reason, Snow.Depth*), and features with

potential collinearity issues (i.e. *Maximum.Temperature* and *Wind.Gust*) were removed. As a result, 34 features spread across 574,709 records were used to predict *Violation.Type* ('citation' or 'warning'). The resulting model had an out-of-bag error rate of approximately 19 percent, indicating a good-enough fit for feature selection. As a result, features with a mean decrease in accuracy less than 15 percent were excluded from any future prediction of a traffic stop's outcome; these features were *Fatal*, *HAZMAT*, *Commercial.Vehicle*, *Work.Zone, Alcohol*, and *Contributed.to.Accident*.

Generalized Boosted Models

Generalized Boosted Models (GBMs) were the first models used to predict a traffic stop's outcome—that is whether the driver received a citation or warning. Applying the full parallelization power of 'caret,' 18 models were easily constructed and evaluated to determine the optimal set of hyperparameters for this particular model. The tuning grid weak-learner ensembles with the following set of hyperparameters: tree sizes of 10, 50, and 100; interaction depths of 1 to 3; and learning rates of 0.01 and 0.1 (the minimum number of observations in a terminal node was held constant at 20). Five-fold cross validation—repeated once—was used to assess the accuracy of each model.

After evaluation, the optimal model—with an accuracy of 73.9 percent—was shown to have 100 trees, an interaction depth of 3, and learning rate of 0.1. A detailed review of all GBMs constructed indicates that while this combination of hyperparameters is optimal, there was little accuracy increase associated with a large forest size or terminal nodes beyond a certain point. For example, a GBM with 50 trees and an interaction depth of 2 had an accuracy of 72.1 percent,

a mere 2% decrease from the optimal set of parameters. The figure below illustrates this principle of diminishing returns:

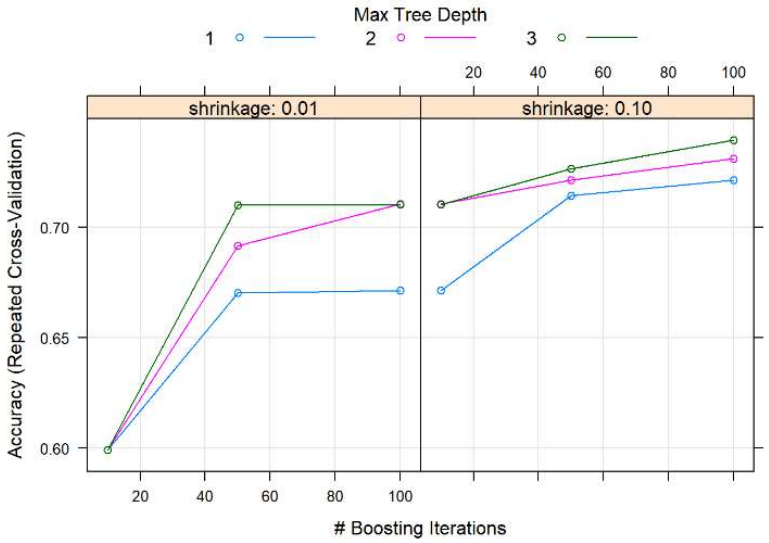

Figure 3. Generalized Boosted Model comparing model fits for various hyperparameters.

As a result, no further GBMs were constructed as the hyperparameters appear to have largely maxed out the capacity of this particular machine learning technique. Using the optimal set of hyperparameters, this GBM was used to predict the outcome of 247,000 traffic stops (40 percent of the original dataset, which was withheld as a test set). The prediction accuracy was 74 percent with an abysmal sensitivity of 47 percent (underestimating the number of citations).

Unfortunately, we suspect that the high-bias, low-variance nature of GBMs resulted in this lower than desired prediction accuracy. In this particular model, a single predictor (*ShortCharge16)* had an importance score of 100 percent while the next closest variables (*AccidentYes* and *MultiInfrTRUE)* had importance scores of just 36 and 32 percent, respectively. As a result, *ShortCharge16* dominated each prediction to the point that the model was unable to appreciate the full complexity of the data and consistently underpredicted 'citation' when *ShortCharge16* was not in the model.

Random Forest (For Citation Prediction)

The decorrelated nature of each tree in the random forest ensemble provided a vast improvement in prediction accuracy over the GBM. To determine the ideal set of hyperparameters for the model, the following variable combinations were evaluated: 2, 4, and 6 features evaluated at each split and forest sizes of 50, 100, and 150. Unfortunately, the 'Caret' package could not be used to parallelize the random forest modelling process due to the large number of categorical variables and, as such, each model had to be constructed in serial, on a single processor, which was a time consuming process that limited the number of hyperparameter combinations that could evaluated. Ultimately, the models generated using the aforementioned hyperparameters generated sufficient variation that an ideal combination of hyperparameters could be selected for final model testing. For this particular dataset, the number of trees in the forest had little effect on the model's accuracy—likely due to the large number of features in the model (27 original features with more than 100 features, including generated dummy variables). The 50-tree ensemble had an Out-of-Bag error rate of 18.0 percent, while the 150-tree ensemble had a 17.6 percent error rate—just a 2.2 percent improvement for a vastly more complex model. The number of features evaluated at each split, on the other hand, had a significant effect on the model's accuracy. An *mtry* of 2 (given a 50-tree ensemble) resulted in an Out-of-Bag error rate of 23.8 percent, while an *mtry* 4 had an 18.0 percent error rate (an *mtry* of 6 had the lowest error rate, but was only a marginal improvement and, as such, was not used to construct the final model). When the ideal model was used to predict the *Violation.Type* (citation or warning) on the test data, it had an incredibly 91.6 percent accuracy. While the model still had a relatively low sensitivity (similar to the GBM) it was still an acceptable 83.9 percent.

To gain a better understanding about the effects of each variable on the model, the package 'randomforestExplainer' provided a detailed report about each feature's role in the ensemble (Paluszynska & et al, 2019). This package greatly improves upon the traditional variable importance measures available in the baseline package.

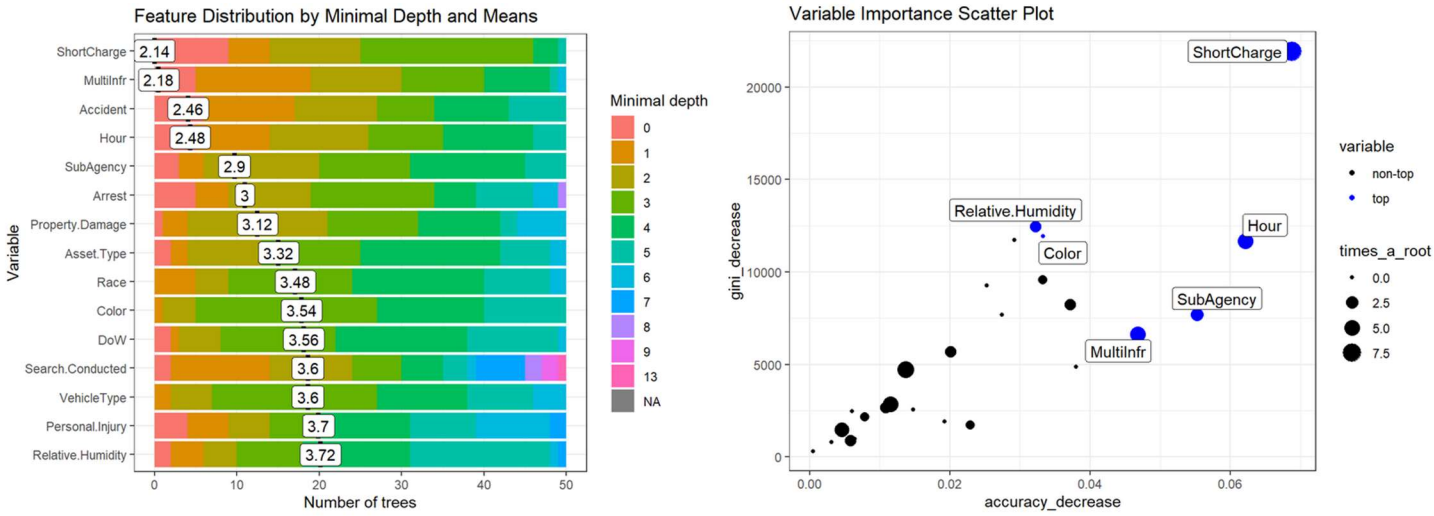

Figure 4. Two standard plots from the randomforestExplainer package that provide a unique look at the importance of each variable considered in the selected random forest ensemble.

Based on the plot's above it is clear that *ShortCharge*, the two-digit number corresponding to the specific line in the Maryland Transportation Code being violated, was by far the most important feature in determining whether a traffic stop resulted in a citation or warning. Other very important features in determining the outcome of a traffic stop were: *MultiInfr*, the variable corresponding to the number of violations committed; *Hour; Subagency;* and *Accident.* Of note, driver demographics (race and gender) did not appear as an important factor in determining the outcome of a traffic stop. Similarly, no evidence in this model suggested that out-of-state drivers received a disproportionate number of citations.

Lasso Penalized Logistic Regression

Since random forest models are a bit of a black box—because it is unclear exactly how any single feature affects the model's ultimate outcome—we applied a lasso penalized logistic

regression to the data to gain a full appreciation of each feature's importance. To generate this model, as was the case with the random forest ensemble, features with extremely low variance were dropped. Additionally, records that contained very low-density levels, such as the vehicle color 'camouflage' and vehicle type 'farm equipment,' were dropped to enable a proper five-fold cross validation of the training data. Consequently, the dataset evaluated by this lasso-penalized logistic regression consisted of 574,237 records spread across 25 features. Using the 'cv.glmnet' package, a lambda of .00175 was identified as the optimal penalty.

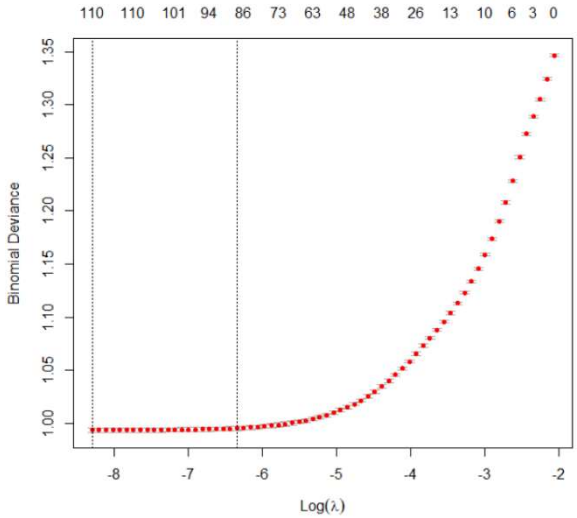

Figure 5. Results of the lasso-penalized logistic regression wherein the optimal value of lambda is identified based on the binomial deviance.

This optimal value occurred one standard deviation from the value of lambda with the minimum binomial deviance, however very little deviance is incurred by dropping 20 percent of the features, so the one standard deviation lambda-value is sufficient for use in future analysis. Evaluating the test data using this model, resulted in a moderate success with an accuracy of 76 percent and a sensitivity of 53.3 percent. While not nearly as accurate as the random forest ensemble, logistic regression provides for a more detailed look at feature importance. Specifically, this model corroborates many of the findings from the random forest ensemble

(such as the importance of *Charge, Accident,* and *Personal Injury* features) while providing the added benefit of a detailed breakout of what charges are most likely to result in a citation.

Feature	Increased Probability of Citation
Charge 16-3 (License Provisions)	.973
Charge 20-1 (Accident Procedures)	.925
Charge 14-1 (Vehicle Theft)	.919
Charge 21-9 (Reckless Driving)	.898
Accident	.895
Personal Injury	.824
Marked Moving Radar	.819
Charge 16-8 (Commercial Licenses)	.818
Property Damage	.792
Arrest	.774

Table 4. Top 10 features with the highest likelihood of receiving a citation compared the baseline model. For 'Charge' the baseline value is Charge 11-1 (General Provisions) while other features are compared against a negative value.

Feature	Increased Probability of Citation
VehicleType—School Bus	.152
Charge 23-1 (Inspection)	.152
Charge 22-2 (Vehicle Equipment)	.165
Charge 11-3 (General Provisions)	.173
Charge 25-2 (Abandoned Vehicle)	.280

Table 5. Five features with the lowest likelihood of receiving a citation compared to the baseline model. Baseline case is the same from Table 2.

The tables above illustrate a trend in which the more serious traffic incidents result in a greater likelihood of getting a citation. For example, it makes sense that an operator would be much more likely (all things equal) to get a citation for reckless driving compared to an expired state inspection. This model does, however, indicate that Hispanic drivers are likely to receive citations at a greater rate than other races—about 58 percent more likely than Asians who were the least cited group. African American drivers also appeared to be cited at slightly greater rate

than average, however statistical significance could not be determined due the limitations of lasso-penalized logistic regression.

Subset Selection

While the previous sections highlight different approaches for variable selection and feature importance generation, is there a better method that can improve interpretability? Can the model be minimized to only include a subset of only variables significantly impacting a traffic stop's outcome? Forward and backward stepwise selection is one such technique. Forward stepwise selection starts with a null model and adds additional features one at a time, while backward stepwise selection is the reverse process. Each technique has its pros and cons.

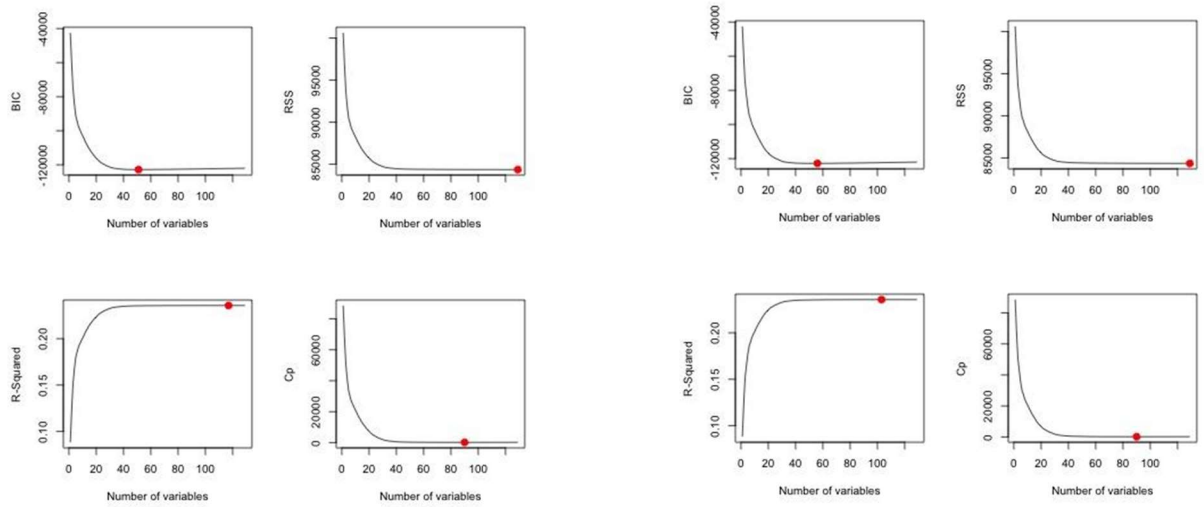

Figure 6. Charts depicting the model accuracies for forward selection (left) and backward selection (right) of the traffic stop data.

Method	BIC
Forward	53
Backward	56

Table 6: Number of Predictors for each method

Based on plots in Figure 6 as well as Table 6 it is observed that BIC minimizes its value for the number of predictor variables in the 50's for both forward and backward stepwise selection whereas the RSS is minimized at 129 predictors (all of them). It is important to

remember that the number of predictors discussed here is not the number of features included in the model, it is the number of predictive criteria. Since the dataset has mostly categorical variables, each possible value is assessed as an individual predictor (one-hot-encoding).

For this analysis, BIC was chosen as the evaluative criterion and the predictors presented above came from 13 variables in the dataset, those corresponding to: accident; personal injury; property damage; search conducted; vehicle type; vehicle color; driver race; driver gender; arrest; type of police asset used; highway; charge; vehicle state; and multi infraction. Ultimately, the inclusion of only these values resulted in an accuracy of 72.9 percent for logistic regression and 73.9 percent for random forest. Specifically, theft (*ShortCharge14*), licensing issues (*ShortCharge16*), and accident scene actions (*ShortCharge20*) had the biggest effect leading to citations. Even though subset selection is not traditionally used for random forests it has the potential to build more efficient trees and prevent overfitting by including unnecessary variables. Unfortunately, in this case, the model generated using lasso logistic regression had a better predictive accuracy and, as such, will be used to draw final conclusions from.

Clustering—Traffic Accident Observations

The objective of this analysis was to identify patterns and similarities among traffic accidents reported in the dataset. There were about 15,500 accidents reported in the data, these were the observations throughout the analytic process. Through previous analyses it was found weather variables did not have much impact or relationship with just about any of the other variables, so only a weather condition variable, with levels of clear and not clear, was included in regard to weather. The rest of the categorical variables from the original traffic violation dataset that were relevant to this clustering were also considered.

The first step of this process was looking at frequency tables for manual removal of a few variables, and some chi squared testing on the rest to narrow down to the final set of variables used in the clustering process. The final set of variables consisted of weather conditions, alcohol, belts, personal injury, vehicle type, vehicle color, race, gender, day type (Weekend or Weekday), and time of day (Morning, Midday, Evening, and Night). The next step was to decide on the clustering method. Hierarchical was selected as k-means is typically not effective with large numbers of categorical variables due to selecting a distance metric chosen and the process for k-means is not as clear with a categorical distance metric. Gower distance was used as the distance metric to measure dissimilarity between the groups. Finally, agglomerative hierarchical clustering was chosen over distinctive hierarchical clustering, for computational cost reasons as well as it being more tailored towards the overall goal.

There were several challenges in the implementation and interpretation of the results, the most important being the decision on the number of clusters and how to interpret clusters with an imbalanced dataset. The number of clusters was decided by running the algorithm with two, three, four, five, six, and seven clusters and selecting the size with the most balanced clusters (in terms of observations per cluster); the conclusion was five clusters. For the interpretation, the

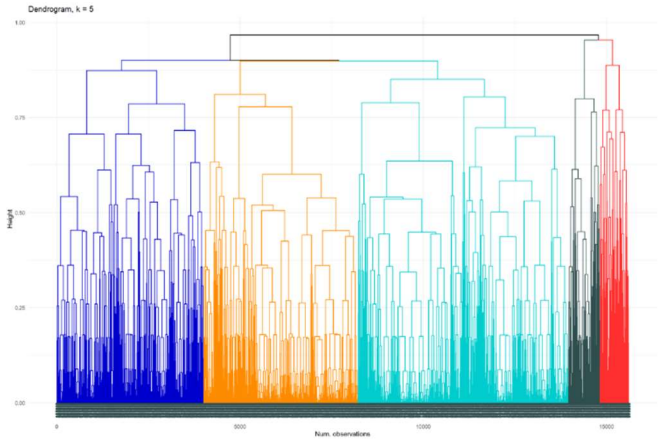

Figure 7. Resulting dendrogram from the hierarchical clustering method to evaluate factors involved with motor vehicle accidents in Montgomery County, Maryland

first step was to look at the dendrogram, pictured in the figure above, but that does not tell much, other than relative cluster size, when over 15,000 observations are involved.

A heatmap that displayed the proportion of observations in each cluster belonging to every level of each factor variable was used to better understand each cluster. This was better, but the imbalance among some of the variables, like alcohol, still made it hard to understand what those proportions really meant. Text was added that displayed the number of observations in the cluster belonging to that factor level, out of the total number of observations belonging to that factor level (the heatmap figure is easier to interpret because the text shows each cell divided by the row sum, and the color shows each cell divided by the column sum). The heatmap is displayed in Figure 2 of Appendix A for reference.

Conclusions

Based on the results above, a few key trends emerged. First, it is apparent that the Montgomery County Police Department principally issues citations based upon the severity of the traffic incident—that is the section of the transportation code being violated, as well as, other incident characteristics (accident, injuries, property damage, etc.). While there is some evidence to suggest that driver demographics (particularly race) may play a small factor, additional analysis is needed to determine the statistical significance of this finding. Another positive sign is that the county's police force generally issues citations at a greater rate for the most serious of traffic offenses—violations like unlicensed (or suspended license) driving, reckless driving, and vehicle theft. These positive findings, indicating that Montgomery County Police Officers apply great judgement while enforcing traffic regulations—a sign that the critical public-police trust should be strong.

Future Studies

Future research should be conducted on the following topics as it will lead to a more complete understanding of how equitable and efficiently Montgomery County Police Department discharge their duties as related to traffic stops:

1. Evaluate the effects of geolocational data (i.e. latitude and longitude) on the outcome of traffic stops.
2. Evaluate the extent to which different races are cited for the same offenses at different rates, whilst other factors are controlled.
3. Assess if the Montgomery County Police Department's efforts to enforce traffic regulations accurately targets the causes of motor vehicle accidents.

References

Paluszynska, Aleksandra, Biecek, Przemyslaw, & Jiang, Yue. (2019). *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*. CRAN. https://cran.r-project.org/web/packages/randomForestExplainer/index.html

Alpert, G., Dunham, R., Stroshine, M., Bennett, K., & MacDonald, J. (2004). *A Report to The National Institute of Justice* (pp. 1–5). U.S. Department of Justice. https://www.ncjrs.gov/pdffiles1/nij/grants/213004.pdf

Baumgartner, F. R., Epp, D. A., Shoub, K., & Love, B. (2017). Targeting young men of color for search and arrest during traffic stops: Evidence from North Carolina, 2002–2013. *Politics, Groups, and Identities*, *5*(1), 107–131. https://doi.org/10.1080/21565503.2016.1160413

Bureau of justice statistics (Bjs)—Traffic stops. (n.d.). Retrieved April 19, 2020, from https://www.bjs.gov/index.cfm?ty=tp&tid=702

Kuhn , M., & et al. (2020). *Classification and Regression Training*. CRAN. https://cran.r-project.org/web/packages/caret/caret.pdf

Liu, C., & Sharma, A. (2019). Are you going to get a ticket or a warning for speeding? An autologistic regression analysis in Burlington, VT. *Transportation Research Interdisciplinary Perspectives*, *1*, 100001. https://doi.org/10.1016/j.trip.2019.100001

U.S. Transportation Secretary Elaine L. Chao Announces Further Decreases in Roadway Fatalities . (2019, October 22). [Text]. NHTSA; United States Department of Transportation. https://www.nhtsa.gov/press-releases/roadway-fatalities-2018-fars

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.)

Table A1. Detailed breakdown of variables in the original datasets (prior to feature generation).

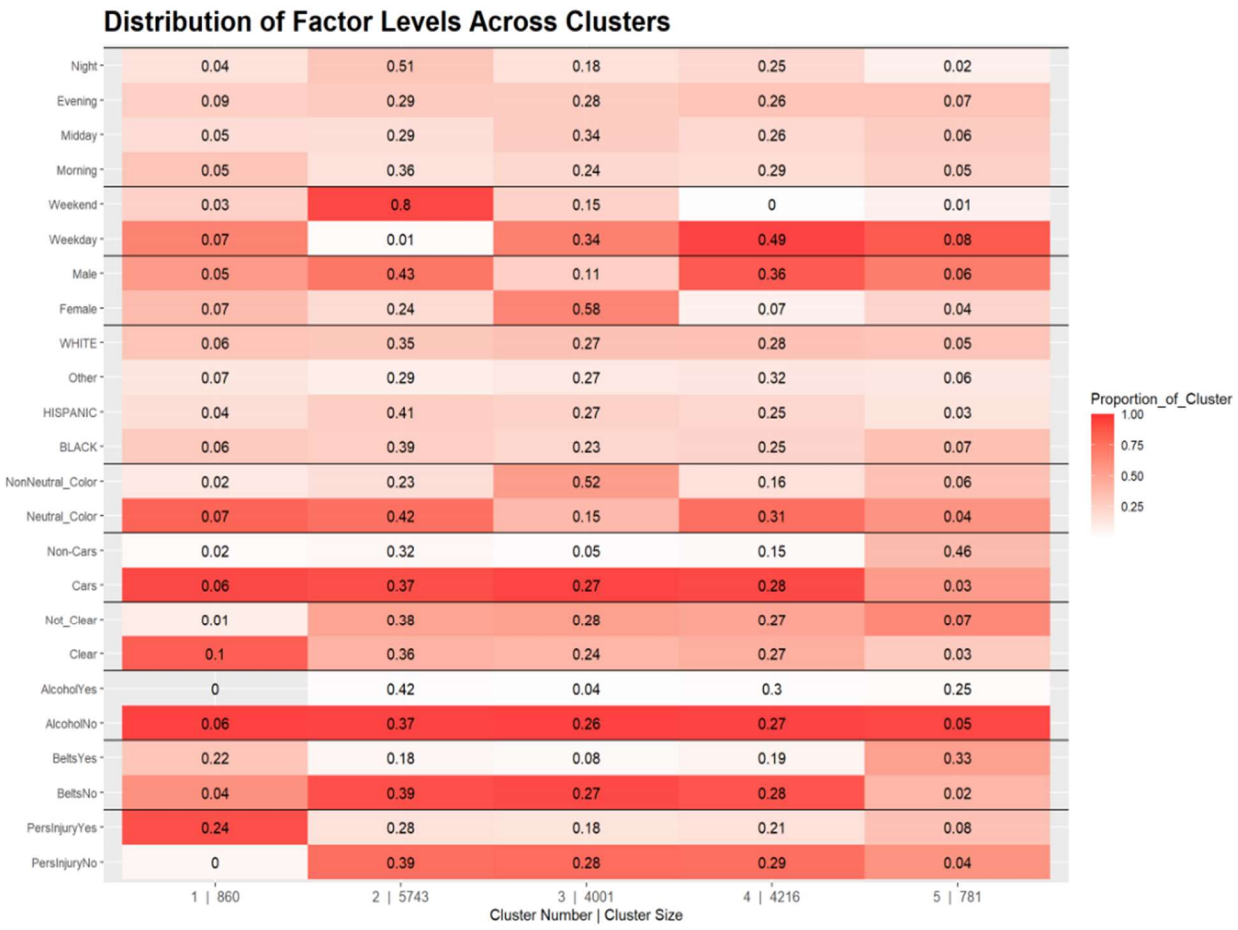

Figure A1. Results from the hierarchical clustering analysis regarding factors impacting traffic accidents.

Appendix B

To review our code or the final dataset, please view our GitHub Repo at:

https://github.com/OutlawSapper/ANLY512_FinalProjectSubmit