Traffic Violations in Montgomery County

Being in the suburb of the nation's capital, Washington DC, and Maryland's business hub Baltimore, Montgomery county is one of the busiest places in the United States. Driving around the county is often stressful, mainly due to the large volume of the traffic which seriously worsen during the office hours. It is very common that drivers swiftly and dangerously change the lanes, angrily honk the preceding cars, and disobey the traffic laws.

On the other side, the state of Maryland is also famous for making a big sum of money from traffic violations, and sadly the large fraction of the money comes from Montgomery county alone. This cost is a bit of frustrating for the poor Montgomery county drivers, but on the other side a smart driver can easily avoid being pulled over or at least manage to receive a warning instead of a citation even after being pulled over. Who are more likely to be the victims of the traffic laws? Can we predict where or not someone gets a citation based on the degree of his violation? Answering these questions is the major goal of this study.

Data Wrangling:

I study the data provided by the Montgomery county government available as comma separated values(CSV). For the ease of analysis, I converted the data in **pandas** *timeseries* DataFrame. The data contains several details of the electronic violations occurring from 2012 to the current date (Nov 1st, 2018).

Data Types:

The data contains 34 columns and 1384236 rows. Each row represents a particular violation. The column contains the categorical features values of a violation. Mainly, there are three types of columns: Categorical, numerical and text. The categorical columns contain values in 'Yes' or 'No'. Numerical column contains floating numbers. The text columns contain description in a form of text.

For data analysis, we treated each type of column differently. The categorical columns were replaced by the dummy variables contains binary number ('0' and '1'). The numeric columns were scaled within the interval [0, 1] using a formula.

$$y = \frac{y - \min(y)}{\max(y) - \min(y)}$$

Where, max(y) and min(y) are respectively the maximum and minimum values of the columns.

The only text column in the data is the *Description* column. For each violation, traffic officers keep a note of the violations as text. Depending on the writing style of police officers, variation in the details, context of the violation and typos, similar violations has been noted in numerous different format. Just as an example, the use of telephone has been described in the following ways.

DRIVER USING HANDS TO USE HANDHELD TELEPHONE WHILEMOTOR VEHICLE IS IN MOTION
DRIVER USING HANDS TO USE HANDHELD TELEPHONE WHILE MOTOR VEHICLE IS IN MOTION
DRIVER USING HANDS TO USE HANDHELD TELEPHONE WHILEMOTOR VEHICLE IS IN MOTION SECONDARY ACTION
DRIVER OF A CLASS H VEH CARRYING PASSENGERS AND INMOTION USING HANDHELD PHONE
HOLDER OF LEARNER'S PERMIT 18 Y/OLDER USING A HANDHELD TELEPHONE WHILE OPER. M/V
.....
.....
HOLDER OF PROVISIONAL LIC 18 Y/OLDER USING A HANDHELD TELEPHONE WHILE OPER. M/V
DRIVER USING HANDS TO USE HANDHELD TELEPHONE
DRIVER USING HANDS TO USE HANDHELD TELEPHONE WHILEMOTOR VEHICLE IS IN MOTION.

Technically, the computer understands the text as a string and even a typo or an extra space between words is considered as a different value. A smarter way to translate the semantic language to the computer's language to convert the text into a mathematical representation. I used word vectorization method to convert the text into an array containing binary numbers. This method first selects the most relevant words, i.e. the words that carry the crucial information. For

example, in the texts above ‘TELEPHONE’ or ‘HANDHELD’ has more significance compared to ‘A’ and ‘IN’. Such high relevant words serve as a word token[1]. Then it converts the text as a linear combination of the tokens.

Missing values:

The data contains some missing values. However, the fraction of data having missing values is small (~18,000) compared to the large data (~ 1 million). We can drop the missing data without a substantial loss of statistical power.

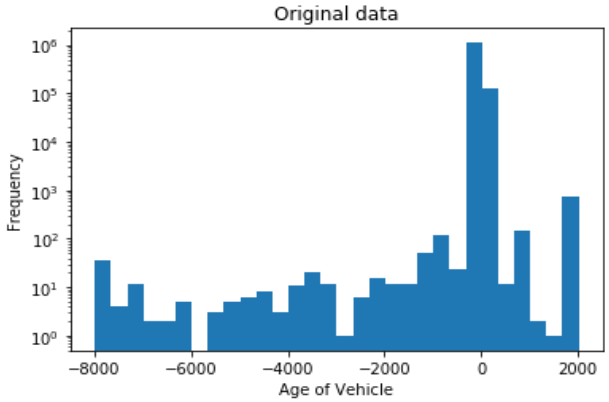


Figure 1: Histogram showing the age of the Vehicles. All the negative values of the data are outliers.

Outliers:

Most of the columns are categorical and it is almost impossible to deal with the outliers in those cases. But the outliers can be corrected for the numerical columns. As an example, a histogram showing distribution of the age of vehicle is shown in Fig. 1. The histogram shows some negative values which is not possible at all. I replaced the outliers with the median age of the vehicle.

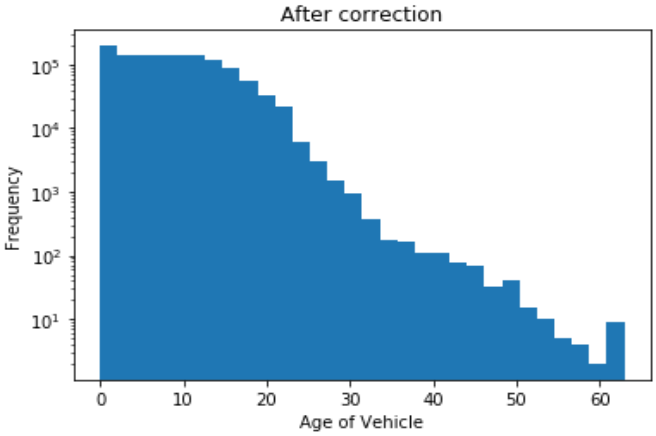


Figure 2: Histogram after outliers’ correction.

[1] In mathematical jargon, it is called basis vectors. The linear combination of basis vector gives the exact translation of the text.

Exploratory Data Analysis:

In average, around 15, 000 violations take place in Montgomery county every month, which amounts to a violation in every fifteen minutes. This number varies with season: The number is in its minimum in winter possibly because people travel much less in the winter. Number of violations increases in the spring and decreases in Summer and Fall. Interestingly, more of the violations occur in late night (after 10 pm) rather than during the busy traffic hours. Among weekdays, the violations are peaked on Tuesday and drop slowly until the end of the week.

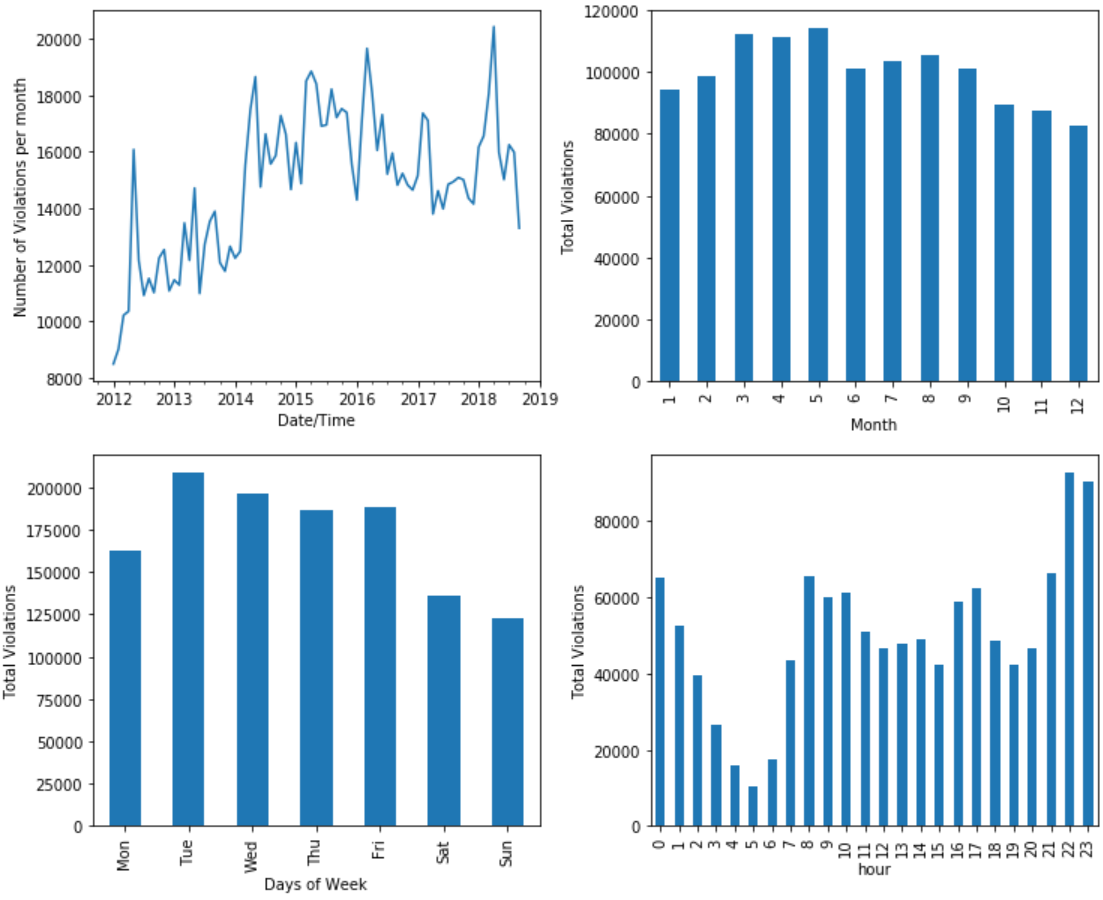


Figure 3: (left top) Number of violations per month showing increasing number of violations over time. (right top) Seasonal variation of number of violations in Montgomery County. More violations occur in spring compared to other seasons. (left bottom) Total violations vs days of week. More violations occur in the weekdays. (right bottom) Total violations vs hours of day. The violations is peaked near the midnight.

As expected, most of the violations occur in the busy highways or streets and there are no major hotspots to avoid. Below is a heatmap of violations occurring at the downtown Silver spring

during rush hours (7-9 am).

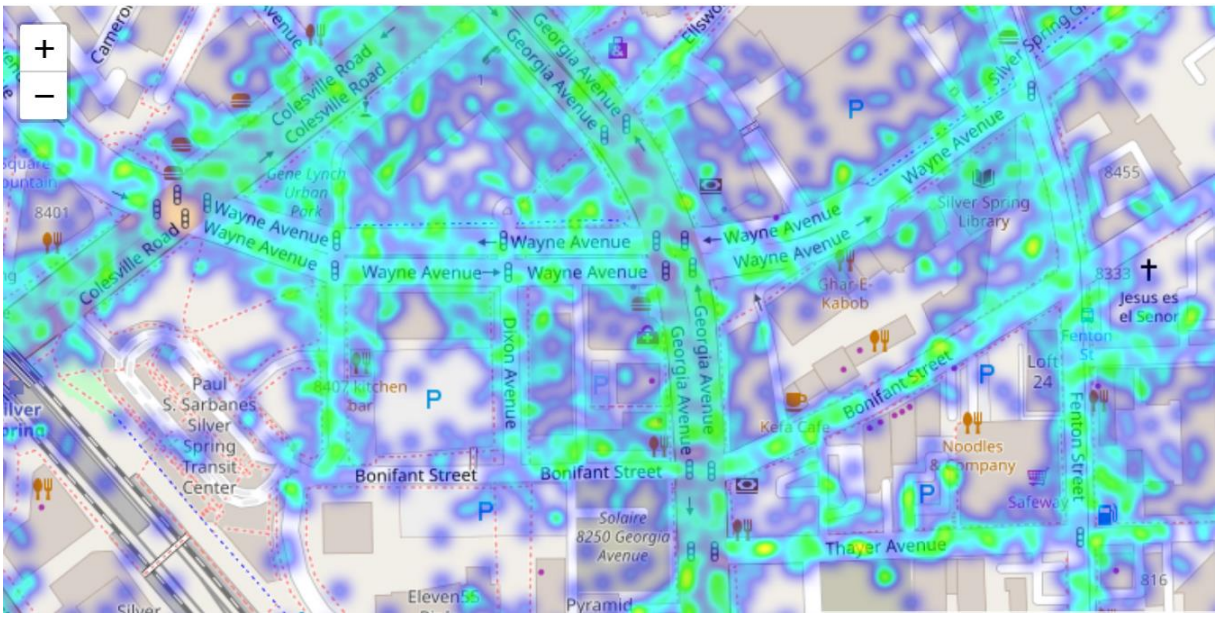


Figure 4: A heatmap showing number of traffic violations in the downtown Silver spring during rush hours (7-9 am)

Driving recommendations:

Below is the list of the most common violation types, which covers a large fraction of violations. The number in left is the number of occurrences the type of violation described as a text. Maximum number of violations take place for the driver's failure to obey the traffic signals, speeding or not having registration card or license in the vehicle. All these can be easily avoided.

108745 DRIVER FAILURE TO OBEY PROPERLY PLACED TRAFFIC CONTROL DEVICE INSTRUCTIONS
94560 FAILURE TO DISPLAY REGISTRATION CARD UPON DEMAND BY POLICE OFFICER
46798 PERSON DRIVING MOTOR VEHICLE ON HIGHWAY OR PUBLIC USE PROPERTY ON SUSPENDED LICENSE AND PRIVILEGE
42870 DRIVER FAILURE TO STOP AT STOP SIGN LINE
42667 DRIVING VEHICLE ON HIGHWAY WITH SUSPENDED REGISTRATION
40831 EXCEEDING THE POSTED SPEED LIMIT OF 40 MPH

Citation vs Warning?

Not all the violations end up to a citation. A lucky driver manages to receive just a warning which gives him/her a great relief from paying an expensive fee or going to a court. In an ideal world, the probability of a driver getting citation or warning should only depend on the degree of violations he/she commits. But in practice, this largely depends on subjective perception of the traffic officer, who similar to every other human being can be biased against the driver's race, gender, or the age of the car, or can be in a bad mood in a particular day. To quantitatively analyze the biasness of the police officers on different features, I define a term citation probability (C_p)

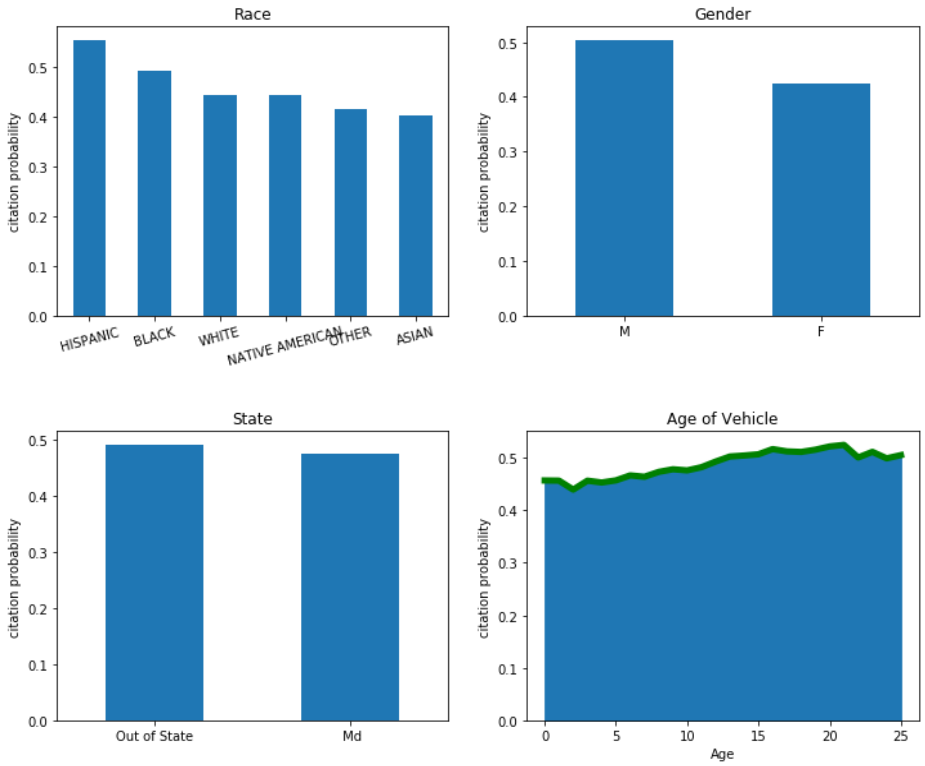


Fig. 5: Comparison of Citation probabilities of different categories. (top left) Among races, Hispanic/Black drivers have more chances of receiving citations compared to Asian and White drivers. (Top right) Male drivers are more likely to receive citations compared to women drivers. (Bottom left) The citation probability does not differ much between the drivers of different states. (Bottom right) The older vehicles are more likely to receive a citation compared to new vehicles.

$$C_p = \frac{N_c}{N_c + N_w}$$

where, N_c and N_w are respectively the numbers of citations and number of warnings received by the drivers of particular subgroup of that category.

Fig. 5 and 6 are the comparisons of citation probabilities of among race, gender, state of driver and the age of Vehicle. Among different races, Hispanic and Black drivers have significantly larger probability of receiving a citation compared to the White and Asian drivers. It is interesting that this has been happening from the beginning and the trend is still on (Fig. 6). Similarly, female drivers are less likely to receive a citation compared to a male driver (Fig. 5). Both of these observations show a significant racial and gender disparity among the drivers, although it is not yet clear how much of that is because of racial/gender bias and how much if any is due to the difference in driving attitudes or other reasons such as age, economic status, and driving history of the drivers.

On the bright side, the good news for Montgomery drivers is that the citation probability is decreasing over time. As shown in figure 6, despite that the number of violations is increasing with time (Fig. 3), the ratio of number of citations to the number of warnings per month is decreasing.

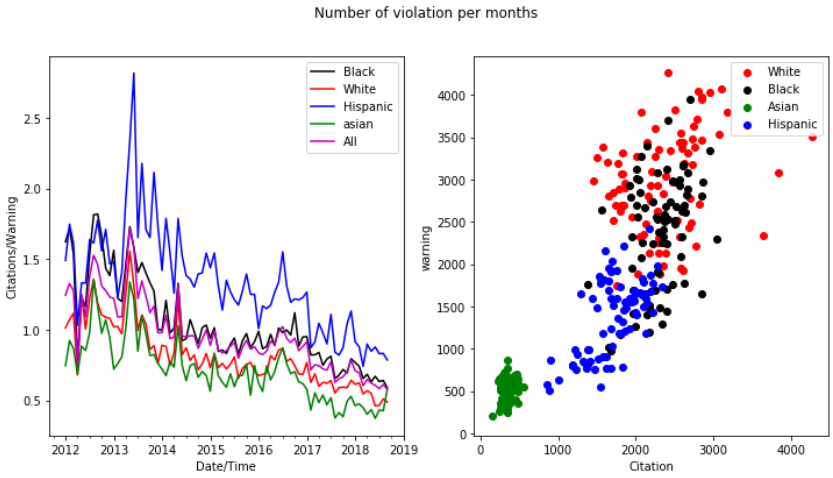


Fig. 6: (left) Variation of number of citations per number of warnings received over time. The number of citations per warnings is decreasing over time. (right) Distribution of number of citations and warnings within different races. This shows that the whites and Asians are the opposite ends of the distribution although the citation probability of this subgroup is significantly lower compared to the drivers of other races.

Women drive less but new:

Driving is still considered as manly work in the US, and there is a preconception that men are better at driving. Further, this stereotype is also supported by a scientific study suggesting that women drive less but are more dangerous drivers. My analysis tells that this is somewhat incorrect. Although women drive much less and much faster compared to men, they actually drive better and follows the rules. They are also less involved in dangerous accidents (even

normalized by their driving population). This is also proven by a smaller fraction of citations that women receive compared to men (Fig. 5).

Another interesting fact is that women appear to get a newer vehicle compared to men. The average age of vehicle driven by male drivers is 9.39 years compared to 7.75 years of average age of vehicle driven by female drivers.

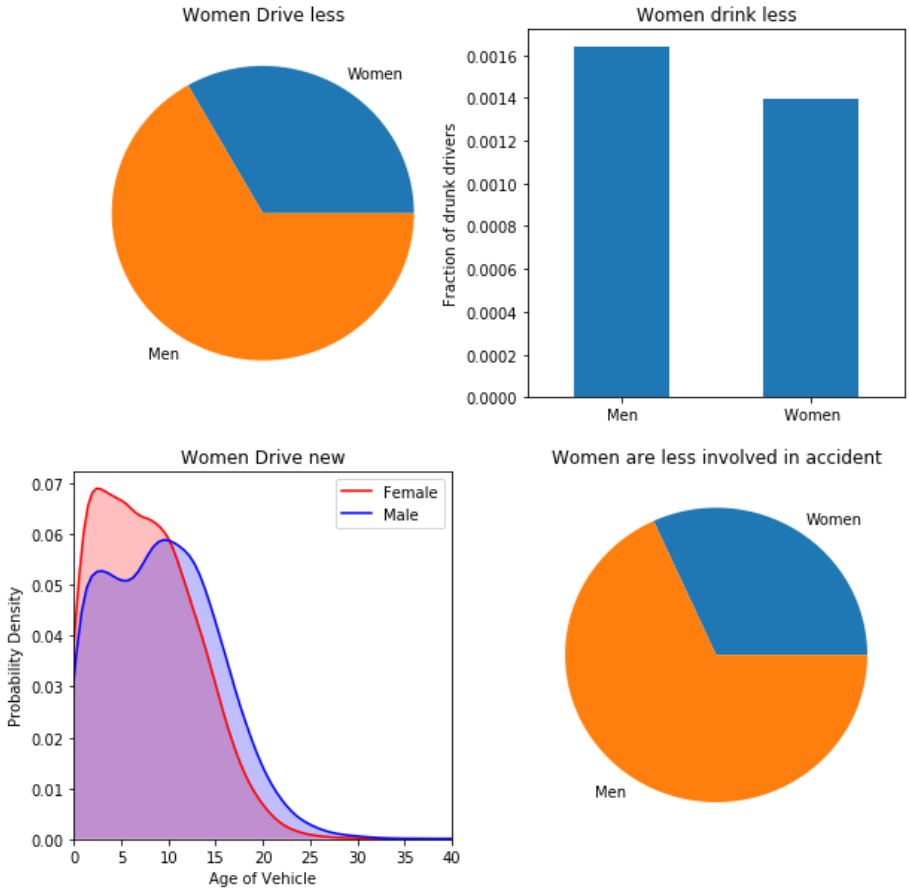


Figure 7: Gender comparison of drivers: (top left) Proportion of violations from male and female drivers. (top right) Fraction of drunk drivers showing larger proportion of male drivers compared to females. (bottom left) Probability distribution of the age of vehicles for male and females. Female drivers get newer vehicle in average. (bottom right) Fraction of accidents caused by male and female e drivers.

Fast Drivers:

Driving around big cities can be sometime like a race, where everyone seems to be competing for a medal. This is particularly bad during rush hours when people require to reach their destination in time. Does this attitude depend on people's race, or Gender or economic status?

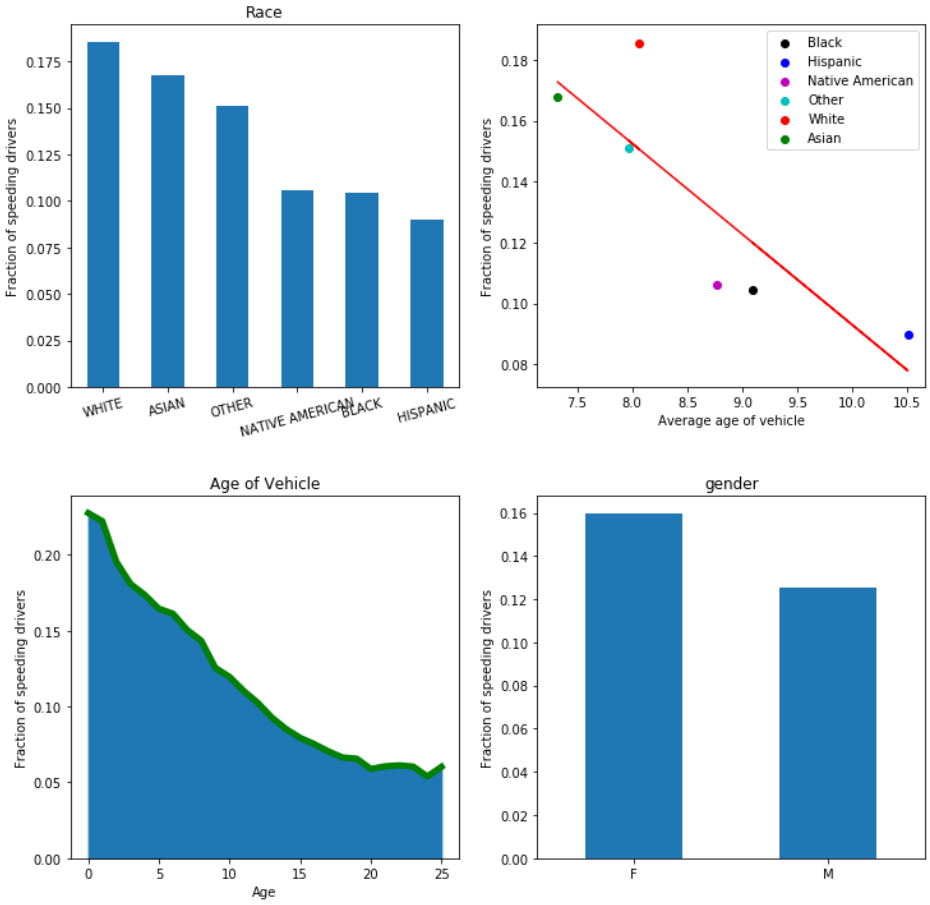


Figure 8: (top left) Fraction of speeding drivers among different races. White and Asian drivers are more involved in speeding compared to the drivers of other races. (top right) Fraction of speeding drivers and average age of the drivers for different races. The variables are not linear indicating that the age of vehicle is not the only factor. (bottom left) Fraction of speeding drivers with the age of vehicle. Drivers with a new car are more likely to speed up compared to the drivers with old cars. (bottom right) Fraction of speeding drivers with Gender. Women are more involved in speeding compared to men.

The data does not tell anything on the speed of the vehicle while being pulled over. To understand whether the driver was pulled over for speeding, I utilized the word vectorization

method. The input text *" EXCEEDING SPEED LIMIT"* is converted to a vector using transformation achieved from the corpus of entire description column. The cosine similarity of the input vector with the description column gives whether or not a particular violation involves speeding. It seems to me that there are number of factors that affects the speeding attitude. First of all, the drivers having new cars are more likely have speeding habits. Similarly, White and Asian drivers generally cause speeding compared to Black and Hispanics. And, women are more likely to cause speeding compared to men.

Since white and Asian drivers drive newer vehicle compared to Hispanic and Black drivers, the difference on speeding attitude among drivers of different races somehow comes from the age of the vehicle they own. But it is not entirely the case: White drivers own older vehicles compared to the Asian drivers in average, but they are more likely to be involved in speeding.

Seatbelt: Neither Drivers nor Police Care:

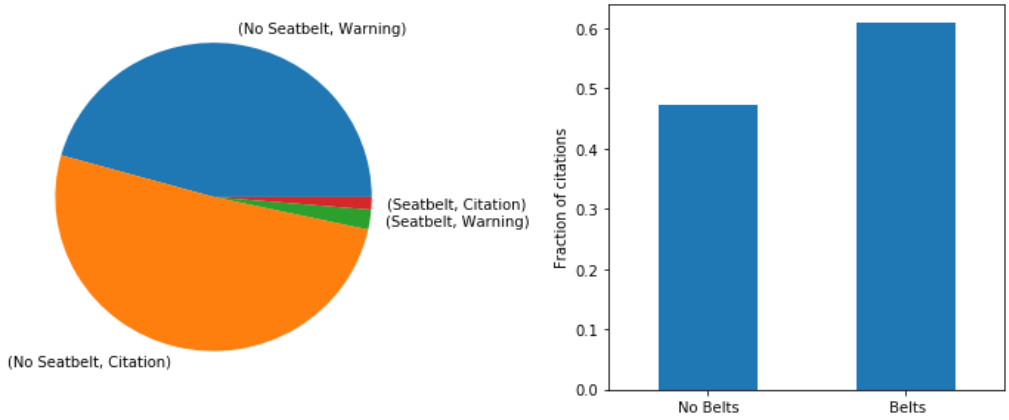


Figure 9: (left) Proportion of seat belt use. Most of the drivers don't seem to use seatbelt. (right) Probability of receiving citation. Drivers wearing belts are no less likely of getting

Everyone knows that seatbelt is for safety. But most of the drivers that were pulled over were found not using seatbelts. This is not so surprising because the drivers that are pulled over are already a violator. However, among the small fraction of drivers using a seatbelt, they are more likely to receive a citation compared to those without a seatbelt. More interestingly, drivers wearing seatbelt receive more citations per warning compared to the drivers who use the seatbelt. This illustrates that the traffic police are indifferent on the use of seatbelt.

Predictive Modelling:

We used supervised learning methods to build a model that predicts whether a driver violating a particular traffic law will get a citation (or a warning equivalently) after being pulled over. We tested several learning methods shown in the table below to achieve the better prediction.

		Precision	Recall	f1-score	ROC Score
Logistic Regression	Citation	0.79	0.69	0.74	0.8503871
	Warning	0.75	0.83	0.79	
Decision Tree Classifier	Citation	0.82	0.68	0.75	0.8614239
	Warning	0.75	0.86	0.8	
Random Forest Classifier	Citation	0.83	0.68	0.75	0.8655828
	Warning	0.75	0.88	0.81	
Multi-Layer Perception	Citation	0.85	0.67	0.75	0.8656781
	Warning	0.74	0.89	0.81	

To make a model, first we separate the data into two sets: The training data that can be used to fit a model. The fit predicts the probability that the driver gets a citation (or warning) and classifies the data into 'citation' and 'warning' categories based on the calculated probability. Then we use the test data to check the performance of our model. To optimize the model, we use randomized grid search method to obtain the best fitting parameters.

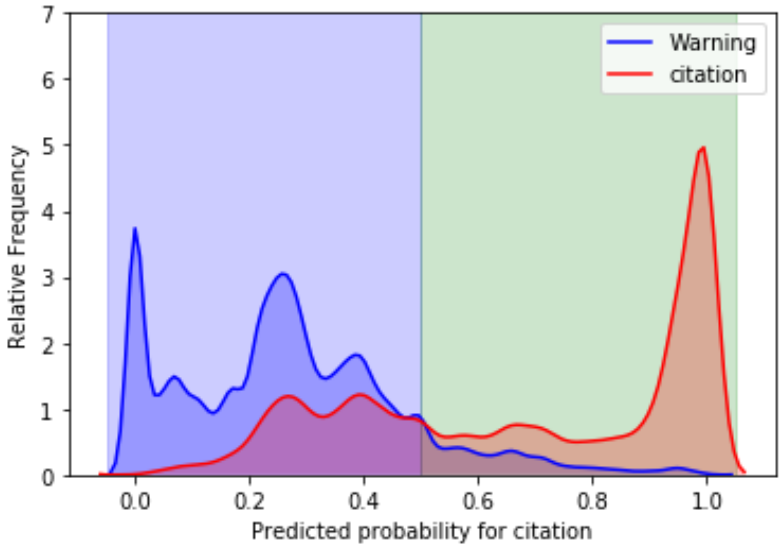


Fig. 10: Calculated probability distribution of test data labelled as Warning (blue) and Citation (red).

The calculated probability distribution is shown in Fig. 10. The red coloured distribution represents the data that is lableed as 'citation', and the 'blue' color represents the subset of data labelled as 'warning'. Our model perform fairly well in predicting the citation probability of

large fraction of data. A comparision of citation proabability is in the Figure below As shown in the table above, all models are almost equally accurate.

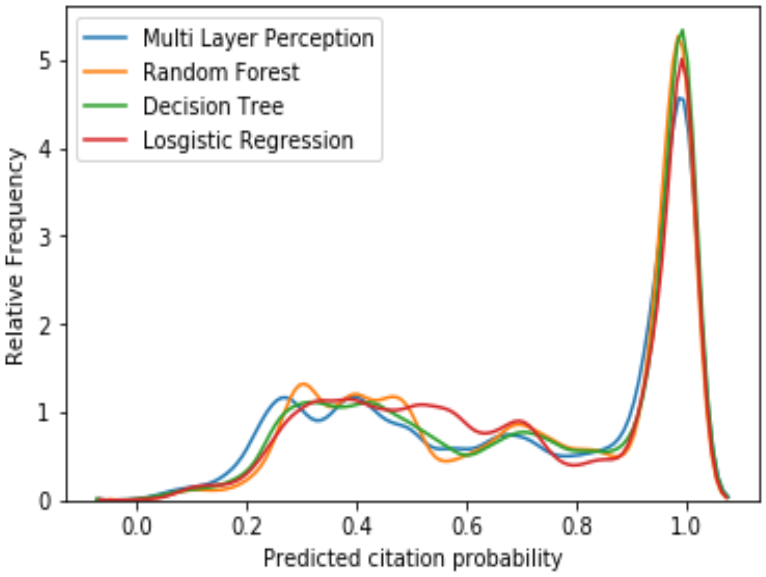
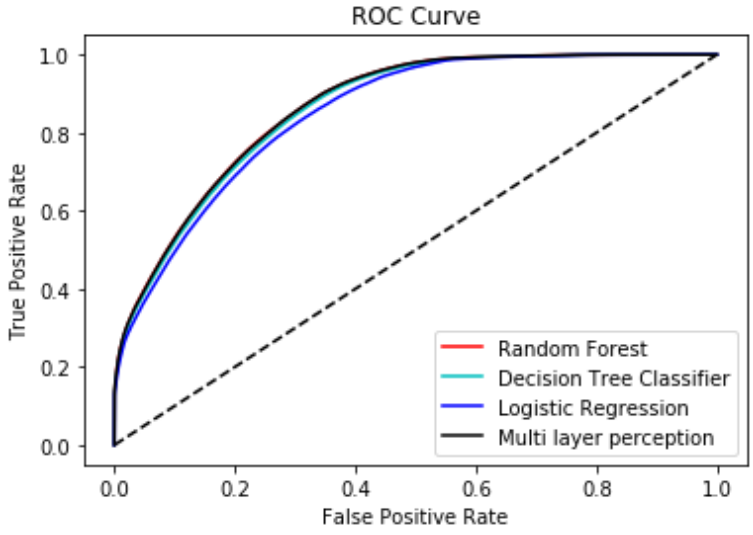


Fig. 11: Calculated citation probability distribution of the data labelled as 'Citation' for various models.

As shown in Fig. 10 and 11, all the models make correct prediction of most of the data, but it also misclassifies some. A better visualization of the fit of the model is the Receiver operating characteristic (ROC) curve. The area under the curve gives the accuracy of the model, which are listed in Table I.



ROC curves for different models.