Traffic Violations in Montgomery County

The state of Maryland makes a lot of money from traffic violations. Reports tell that the state profits almost around 62 millions in a year from speeding camera alone. The major fraction of this income comes from the Montgomery county. For a Montgomery driver, knowing the chances of getting traffic citations is a very useful – both for curiosity and for avoiding the traffic citations. In this report, I will summarize the interesting trends of traffic violations in Montgomery county over the past five years. I present a statistical model that predicts the probability of receiving 'citation' after getting pulled over.

Data Wrangling:

I used the data provided by the Montgomery county government. The data is 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 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. Columns contain the values of violations. There are three types of columns: First are the categorical columns containing 'Yes' or 'No' values. Secondly, some columns are in floating numbers and thirdly there are Text columns that contains values as a text.

For data analysis, I treated each type of column differently. I replaced the categorical columns by the dummy variables contains binary numbers ('0' and '1'). For numerical columns, I scaled them using to range between '0' and '1'

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

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

Natural Language Processing:

The data contain ‘Description’ column which is a record of traffic officers describing the violations in a form of text. Grouping the data by the description column identifies 12000 unique violations. However, the problem with this method is that the computer program categorizes by the string of the text but not by its literal meaning. Just as an example, there are several unique values of description column that are literally identical.

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.

Thus, even a typo or an extra space between words makes a different category. To translate this column into a programming language, I used *word vectorization* method. This method first selects the most relevant words, *i.e*. the words that carry the most 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*. Then it converts the text into a numeric array that represents the presence (or absence) of a word token.

Missing values: The data contains some missing values. The fraction of data having missing values is small (~18,000) compared to the large data (~ 1 million). I dropped the missing data without a substantial loss of statistical power.

Outliers:

Most of the columns are categorical and it is almost impossible to deal with the outliers in those cases. In numerical columns, I corrected the outliers by replacing with their median value. As an example, Fig. 1 shows a histogram of the age of vehicles. It contains some negative values which does not make sense. After replacement, the histogram looks much cleaner (Fig. 2).

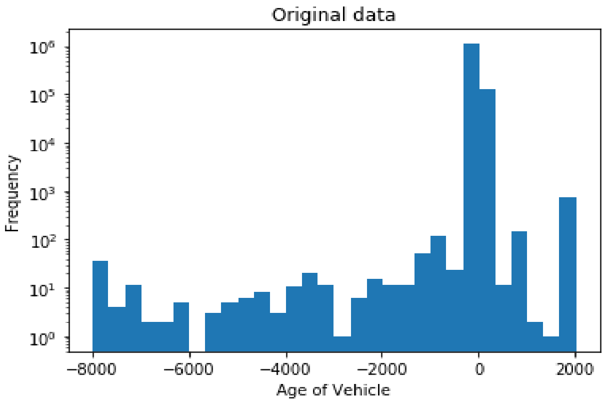


Figure 1: Histogram showing the age of the Vehicles. All the negative values of the data are outliers.

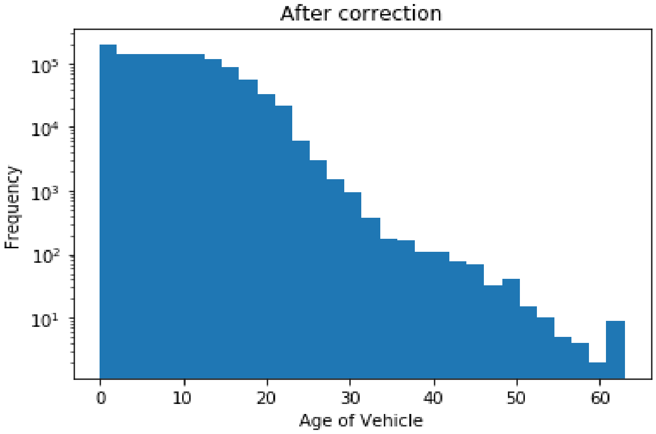


Figure 2: Histogram after outliers' correction.

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. Violations are peaked in the Spring and decrease in Summer and Fall. More violations occur in late night (after 10 pm) compared to the rush hours. Among the weekdays, violations are peaked on Tuesday and drop slowly until the end of the week.

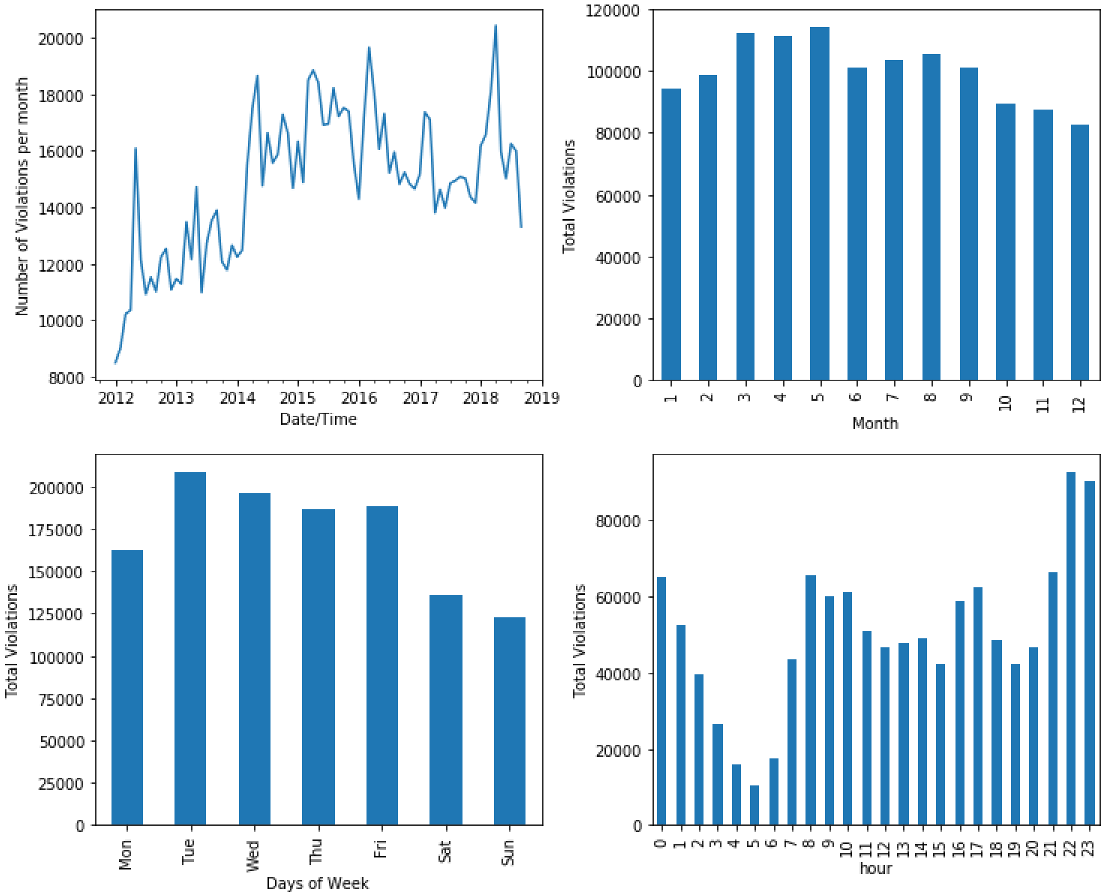


Figure 3: (left top) *Number of violations per month showing increasing number of violations over time. (right top) Seasonal variation of number of violations in Montgomery County. More violations occur in spring compared to other seasons. (left bottom) Total violations vs days of week. More violations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

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).

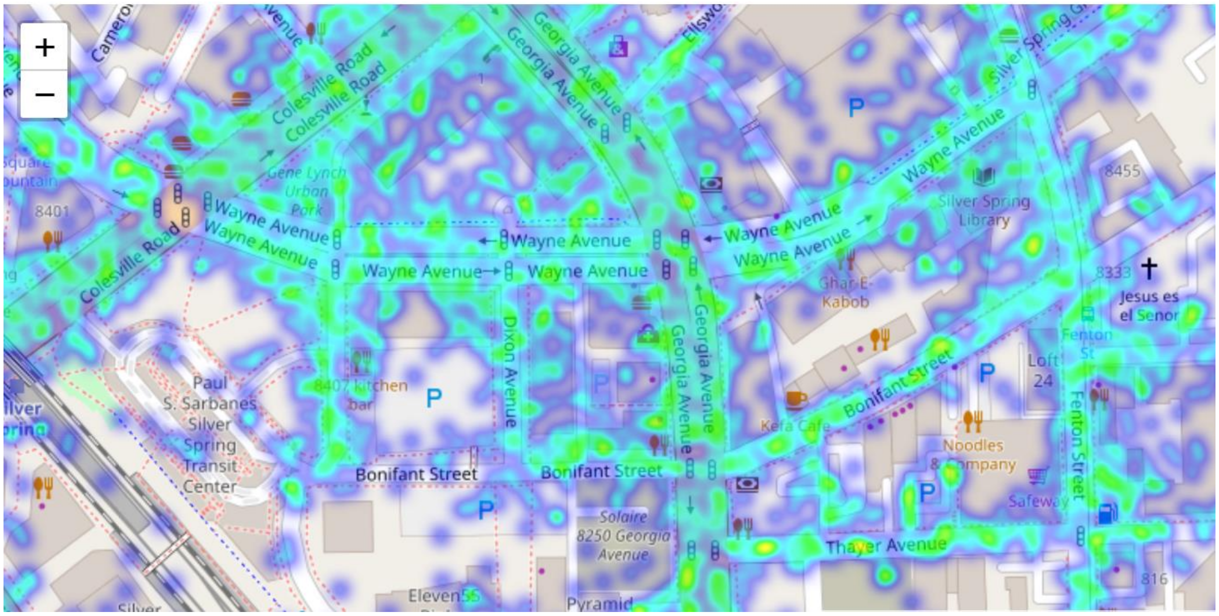


Figure 4: A heatmap showing number of traffic violations in the downtown Silver spring during rush hours (7-9 am)

Most Common Violations:

Below is the list of the most common violation types, which covers a large fraction of violations. The number in left is the occurrence of the type of violation described as a text.

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 – Some receive just a warning. In an ideal world, the probability of a driver getting citation or warning should only depend on the degree of violations the driver commits. But in practice, this largely depends on subjective perception of the traffic officer, who 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)

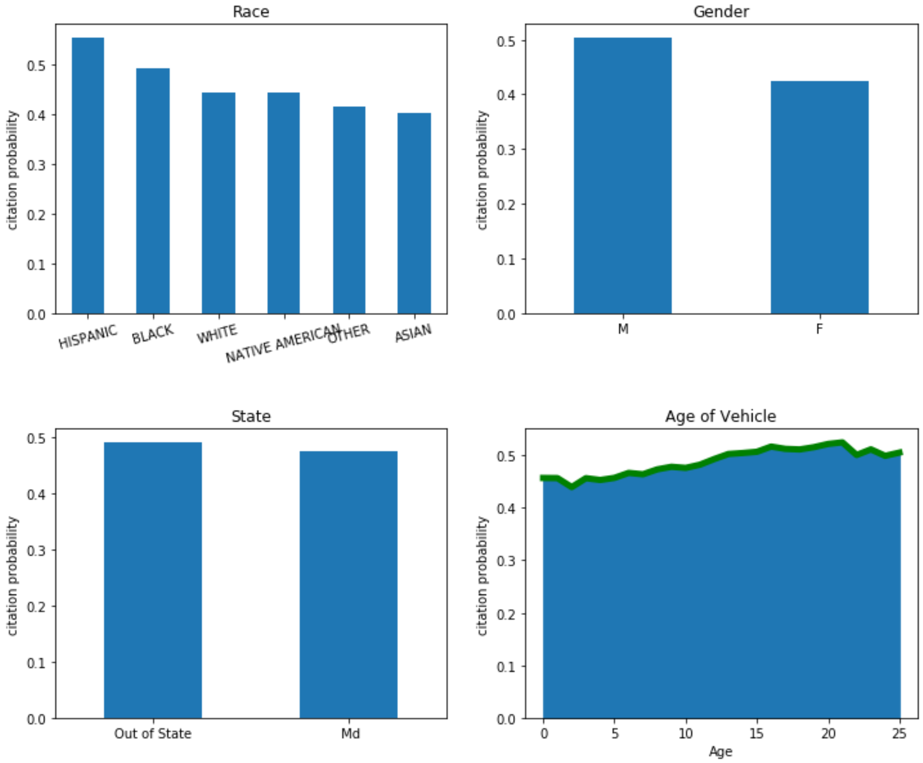


Fig. 5: Comparison of Citation probabilities of different categories. (top left) Among races, Hispanic/Black drivers have more chances of receiving citations compared to Asian and White drivers. (Top right) Male drivers are more likely to receive citations compared to women drivers. (Bottom left) The 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. Similarly, female drivers are less likely to receive a citation compared to a male driver. Both of these observations show a significant racial and gender disparity among the drivers. Although it is not yet clear from the data if this is entirely a racial/gender bias or there are differences in driving attitude or other reasons such as age, economic status, and driving history of the drivers.

The good news for the 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.

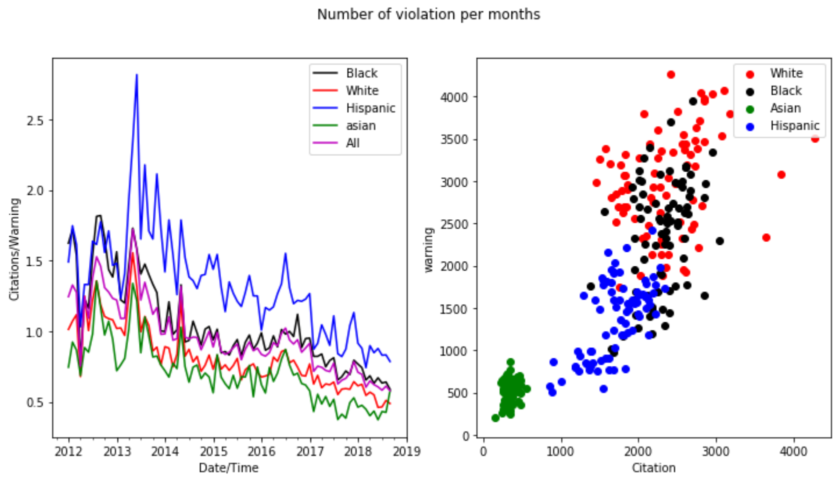


Fig. 6: (left) Variation of number of citations per number of warnings received over time. The number of citations per warnings is decreasing over time. (right) Distribution of number of citations and warnings within different races. This shows that the whites and Asians are the opposite ends of the 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, this stereotype is also supported by a scientific study. My analysis tells that this is somewhat incorrect. Although women drive much less and much faster compared to men, they drive better and follows the rules. They are also less involved in dangerous accidents (even normalized by their driving population). This is also verified by a smaller fraction of citations that women receive compared to men.

Another interesting fact is that women in average drive 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

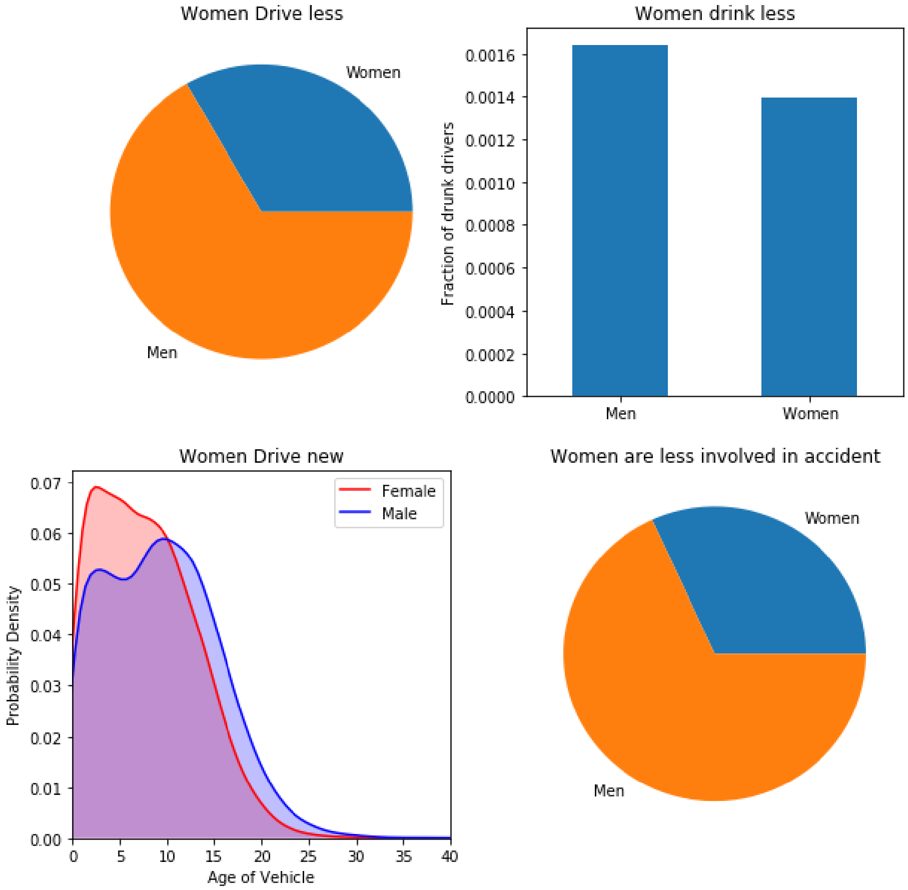


Figure 7: Gender comparison of drivers: (top left) Proportion of violations from male and female drivers. (top right) Fraction of drunk drivers showing larger proportion of male drivers compared to females. (bottom left) Probability distribution of the age of vehicles for male and females. Female drivers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

age of vehicle driven by female drivers.

Fast Drivers:

To understand whether the driver was pulled over for speeding, I utilized the word vectorization method. I converted an input text “*EXCEEDING SPEED LIMIT"* 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.

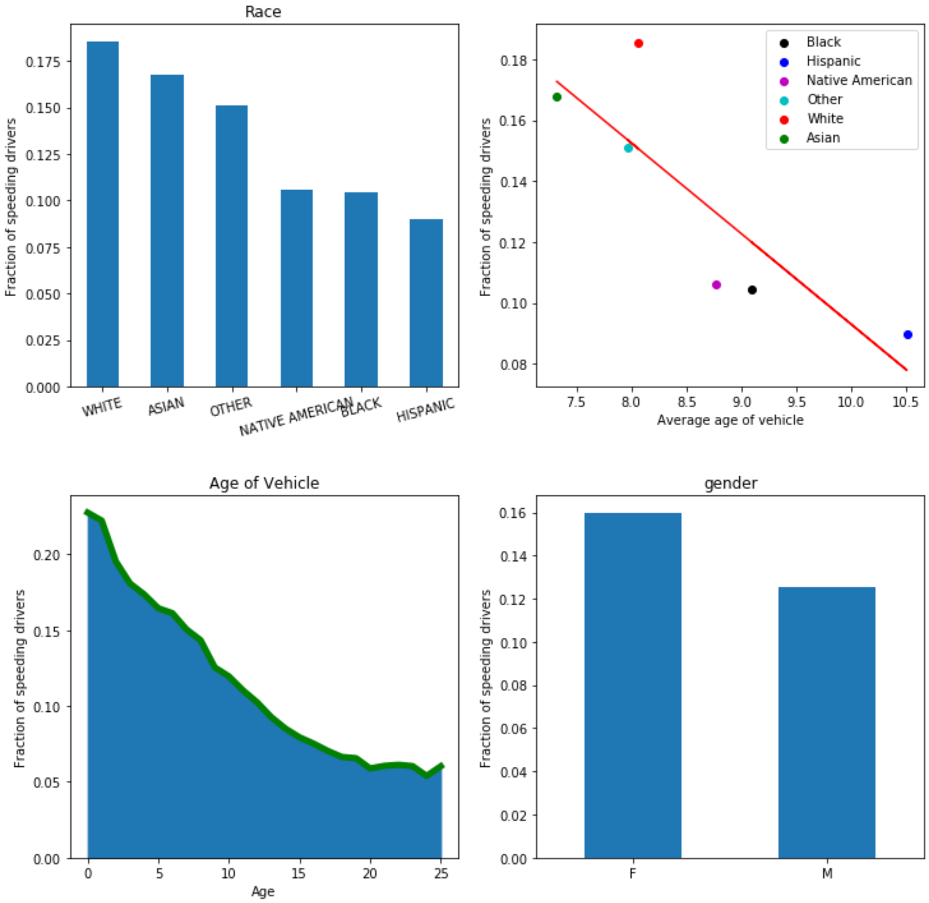


Figure 8: (top left) Fraction of speeding drivers among different races. White and Asian drivers are more involved in speeding compared to the drivers of other races. (top right) Fraction of speeding drivers and average age of the drivers for different races. The variables are not linear indicating that 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.

There are a number of factors that affects the speeding attitude: Drivers having new cars are more likely to have speeding habits. White and Asian drivers are more involved in speeding compared to Black and Hispanics drivers. And, women are more likely to cause speeding compared to men.

Seatbelt: Neither Drivers nor Police Care:

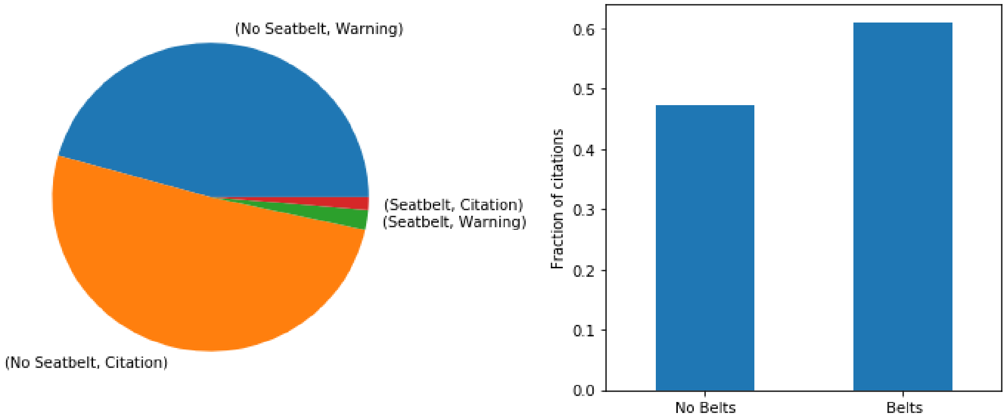


Figure 9: (left) Proportion of seat belt use. Most of the drivers don't seem to use seatbelt. (right) Probability of receiving citation. Drivers wearing belts are no less likely of getting

Most of the drivers that were pulled over were found not using seatbelts. Among the small fraction of drivers using a seatbelt, they are more likely to receive a citation compared to those without a seatbelt. This illustrates that the traffic police are indifferent on the use of seatbelt.

		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	

Predictive Modelling:

I used supervised learning to build a model that predicts whether a driver violating a particular traffic law gets a citation (or a warning equivalently) after being pulled over. I tested several learning methods to achieve a better prediction. To optimize the model, I used *randomized grid search* method. As shown in the table above, all models are perform well and are equally accurate.

The calculated probability distribution of a model using Random Forest classifier is shown in Fig. 10. The red colored distribution represents the data originally lableed as 'citation'. The 'blue' color represents the subset of data labelled as 'warning'. My model perform fairly well in predicting the citation probability of most of the data.

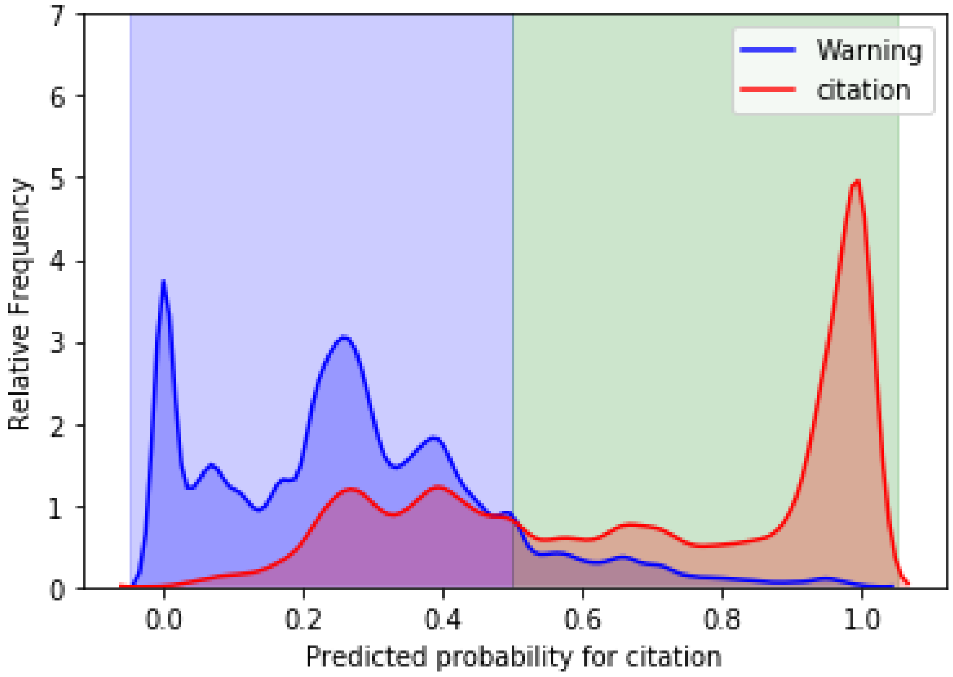


Fig. 10: Calculated probability distribution of test data labelled as Warning (blue) and Citation (red).

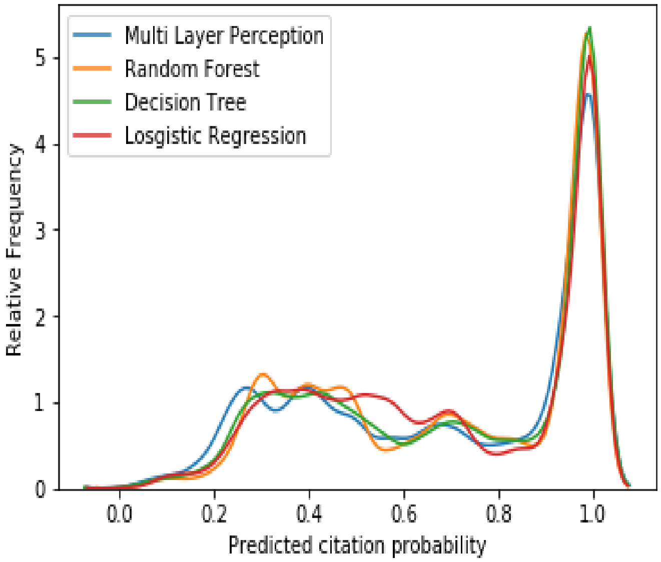
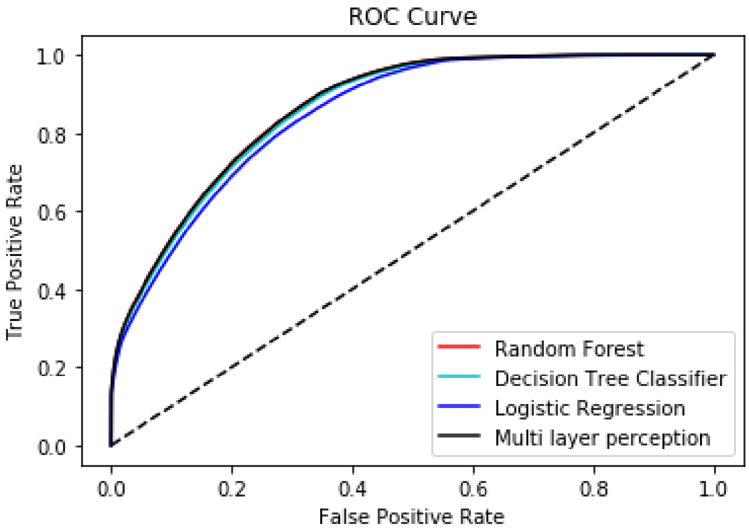


Fig. 11: Calculated citation probability distribution of the data labelled as 'Citation' for various models.

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.