MSC DATA SCIENCE & ANALYTICS DISSERTATION

Mesa City Traffic Violation Prediction

Author:

Shashank Sanjee Venkata Chalapathi

Supervisor:

Dr. Niamh Cahill

Maynooth University

Department of Mathematics & Statistics

MAYNOOTH UNIVERSITY

Department of Mathematics & Statistics

Master's in data science and analytics

Mesa City Traffic Violation

by Shashank Sanjee Venkata Chalapathi

Abstract

This research describes a study of traffic stops in Mesa, Arizona, in an attempt to recognize the racial inequalities and predicting arrest violations at different locations. In this work, Mesa stop data for the 4 years is analysed using different data-processing approaches. From my results one can notice that police taking decisions regarding stops are racially biased and the most common violation in the city is "Speeding". In addition, there are some correlations between number of cases and hour of the day. This paper investigates machine-learning-based violation prediction at different locations. Multinomial model, Decision Tree, Random Forest, Neural Net, and K nearest neighbours models were implemented and a crime prediction accuracy between 34% to 83% were obtained when predicting different arrest violations in Mesa. Analysis and prevention of violations is a systematic approach for identifying and analysing crime patterns and trends. Our system can predict regions with high probability of arrest violations and can visualize areas that are prone to violation and can help to prevent these violations before happening by increasing rules and detection devices. In addition, I have incorporated an interactive R Shiny application that showcase the output from the model.

Keywords: Multinomial model, Decision Trees, Neural Net, Random Forest, K nearest neighbours, Machine Learning.

TABLE OF CONTENTS

LIST OF TABLES, GRAPHS, MAPS AND EQUATIONS

1 BACKGROUND/INTRODUCTION:

Every year, more than 20 million Americans are pulled for traffic violations, making this one of the public's most common ways of communicating with police (Department of Computer Science, Management Science, Communication, Applied statistics, Stanford, CA, USA). Among these police stops, Black Americans are approximately 2.5 times more likely to be arrested each year than white Americans (Starr, 2016). Racism has existed in the United States since the colonial era when white Americans were granted certain extra privileges and freedoms where other races and religions were not provided the same benefits.

To examine these ethnic inequalities with police and the public, The Stanford Open Police Project has compiled and consolidated more than 100 million traffic stop and search data reports nationwide by using the Police-Public Contact Survey (PPCS offers comprehensive details on the characteristics of people who have had some kind of police contact during the year, including those who have called the police to report a crime or have been pulled in a traffic stop (Elizabeth, et al.)) (Pierson et al., 2020). They collected, analysed, and released a data collection that documented nearly 100 million traffic arrests conducted over nearly a decade by 21 Public patrol agencies and 35 local police agencies.

Mesa is the 25th most congested city in United States and 20th among metropolitan cities (Traffic Patterns in Phoenix Arizona). Due to high population in the city, there is more traffic congestion. So, from the collection of datasets, I have chosen the Mesa city (AZ state) dataset for my research which has 96621 stops and 18 variables to explore.

There are proofs of bias against both Black and Hispanic car/truck drivers while stopping and arresting (Pierson et al., 2020 and Goel et al., 2017). So, I examined the trend of stopping different people and arresting them for different

violations, in my first analysis. The aim of the analysis is to explore interactions between police and people, i.e. whether police officers are biased against an ethnicity and also to find the trend in number of crimes committed over the years either increasing or decreasing.

On exploring further, I found out that peak hour traffic in Mesa city is around 7am to 9am and 5pm to 6pm (Traffic Patterns in Phoenix Arizona). Hence from the data, to analyse this peak hour traffic and number of cases during these hours across Mesa city, in my next analysis I sorted the data into total number of cases w.r.t. hour of day and explored the traffic peak hours in the city

Weekend traffic around the city is less compared to weekdays except on year's busiest travel days, like Christmas, Fridays on summer etc (Cristiano, Jason). So, to analyse this I sorted the data into total number of cases w.r.t. day of week.

Next, to analyse popular violation in the city, I found out that most common violation in the city is Speeding(Traffic Patterns in Phoenix Arizona), to verify this from the data obtained, I organised the data into total number of cases for each violation and the results seems to be same.

Instead of focusing on causes of violation such as driver history or motive for committing the violation, my research focuses primarily on types of violation happening every day. This will predict the arrest violation group that has a greater likelihood of occurrence across Mesa city. The study depends on a different variable such as sex, location, race, age, etc.

For certain areas police are now using spatial crime models to minimize these crimes. The Los Angeles Police Department , for example, used spatial crime forecasts to pre-emptively assign patrol units and found that regional criminal intelligence decreased violent crimes by 5.4 percent and homicides by 22.6 percent (Uchida et al., 2012). So, to analyse arrest violation hotspot around

Mesa city different machine learning models are used for estimating the percentage of different arrest violations and assessed the results in this paper.

The rest of this paper is organized as follows: first, I have provided a description of traffic stop data, which includes details about data, missing data, and traffic stop characteristics held in Mesa city and drivers characteristics who has been stopped along with some tables and graphs. Followed by preliminary data analysis. Next, I have provided a detailed overview of all the statistical and machine learning methods used for the visualizations, modelling, and development of the shiny app. Then I have displayed the results for all my analysis and results from model and along with the predictions from model for different locations. This is followed by a discussion of the results, what is the use of this analysis and what led me to perform this analysis. Finally, I present the conclusion of the paper with suggestions for future work.

2 DATA DESCRIPTION

***Overview*:**

This segment outlines the stops for the period from 1 January 2014 – 31 December 2017 based on traffic stop data collected by PPCS. This section is divided into 4 parts reporting: 1) Table of data description 2) missing or invalid traffic stop data, 3) traffic stop characteristics conducted between 2014-2017, and 4) vehicle and driver characteristics stopped by police officers.

The first section includes a description of the data in Table 2.1 and the percentage of missing data for the variables used in this analysis in the Table 2.2. The third section contains Tables 2.3 – 2.4, which detail traffic stop features around the city for 2014-2017. Table 2.3 offers a monthly breakdown of the 2014 to 2017 traffic stops in city. The fourth section contains Tables 2.4, which detail driver characteristics (e.g., age, gender, race / ethnicity) that were stopped in 2014-2017 by police officers.

2.1 Summary of Data:

Stanford 's open policing project reported 96,621 citizens' stops for Mesa City between 2014-2017. The number of stops and arrests for different racial groups differed significantly depending on the race, age, and violation committed by the drivers. Despite these constraints, there is no statistical evidence of the initial decision to stop traffic. Accordingly, it is possible to draw city-wide conclusions about whether there are racial / ethnic differences in stopping behaviour. In my study I focused on whether there were racial / ethnic differences for traffic violations and the outcome of those violations.

Only officer-initiated traffic stops should be used to better investigate problems concerning racial / ethnic inequalities. It is important that for each stop, only one entry should be used in the data, so that stop information is not

duplicated (i.e., stops that are inserted multiple times into the dataset are removed).

Each row in the data represents a stop for a violation committed by a driver. For each stop, there are entries as in date and time of the violation committed, the location where it was committed, a person's sex, age, race, and type of violation that one has committed and outcome for that violation.

Table 2.1 shows the description of each variable and sample values for the variables.

Table 2.1. Data Description table

Variable name	*Values*	*Explanation*
date	*01/01/2014*	*Date and Time when the violation took place.*
time	*05:04, 12:33, 13:52* *Time (24-hr format)*	
location	*700 E FRANKLIN AVE, MESA, etc.*	*Address where the violation took place*
lat	*33.40238*	*Co-ordinate values for the location*
lng	*-111.8161*	
subject_age	*36, 42, etc.*	*Age of the person who committed the violation.*
subject_race	*Hispanic, white, black, other, unknown, Asian/pacific islander*	*Race of the person who committed the violation.*
subject_sex	*female, male*	*Gender of the Person who committed violation*
officer_id_hash	*d4eed1c4dc, e24408e0fd, 4ab12ca221 etc.*	*Police officer ID*

type	*vehicular, pedestrian*	*Violation type whether the person was driving a vehicle or was pedestrian.*
violation	*NOISE AFTER 10 PM TO 6 AM, NO PROOF OF INSURANCE, SPEED NOT R&P/FTC SPEED TO AVOID A COLLISION, SUSP/REVOKED LIC, etc.*	*It tells us the exact violation; a person has committed.*
outcome	*citation, arrest*	*It gives the outcome of the violation.*

2.2 Missing Data:

From the data, For the reasons noted, the following traffic stop data have been removed from further analysis:

- 7309 non-traffic cases were removed.
- 7092 cases with subject race as unknown were removed because for examining issues surrounding racial/ethnic disparities, subject race is must.
- 2203 contacts with missing data on stop time(n=190), stop location(n=1217), driver's age(n=96), driver's sex(n=19), violation(n=42) were removed because it is not possible to predict with all of these values missing.

Table 2.2: Missing Data Analysis from all 2014-2017 Traffic Stops

Variable name	*Percentage missing*	*Total number of Valid cases*
date	0.00	83673
time	0.18	83523
subject_age	0.11	83577

subject_race	0.00	83673
subject_sex	0.02	83654
violation	0.00	83673
arrest made	0.00	83673
lat	1.45	82456
lng	1.45	82456
type	0.00	83673

For the 96,621 traffic stops, Table 2.2 shows the percentage for missing data for each variable used in analysis. The first column shows the variable names, in the second column percentage of missing data, and the number of valid cases left. Among the variables available in the data set, the variable with the highest percentage among missing data is violation type (5.65%), followed by geo-coordinates (1.53%). The remaining variables have less than 1.0 per cent of missing data.

The remaining analyses in this study are therefore focused on 82199 driver traffic stops performed during 2014-2017, which represents 85 per cent of the total stops reported across 4 years.

2.3 Traffic stop characteristics

2.3.1 Traffic Stops Month-wise:

Table 2.3 shows the short-term number of traffic stops by month for years 2014 through 2017. At city level March accounted for the largest number of stops (11.68%), followed by January (11.01%), February (10.78%), August (9.42%). The lowest percentage of the city's traffic delays occurred in November (6.42%), October (6.61%), June (7.6%), and April (7.09%). However, average city-level stop activity is relatively constant over months, with a gap of just 3 percent between the slow and busy months .

Table 2.3: Traffic Stops by Month

Months	Total Cases	Percentage
Jan	9053	11.01
Feb	8867	10.79
Mar	9603	11.68
Apr	5832	7.09
May	6076	7.39
Jun	5816	7.08
Jul	6061	7.37
Aug	7749	9.43
Sep	6440	7.83
Oct	5439	6.62
Nov	5279	6.42
Dec	5984	7.28

2.4 Driver Characteristics by Race:

Table 2.4 details the driver characteristics that police officers stopped between 2014-2017, including the percentage of racial groups (e.g. White, Black, Hispanic, Asian / Pacific Islander, and others).

Table 2.4: Total Traffic Stops by Population Percentage

Driver's race	Total Stops	Population	Percentage Total stops
Asian/pacific islander	1198	45452	2.64

black	5781	66911	8.64
Hispanic	15235	540866	2.82
other	1918	374025	0.51
white	58067	1181965	4.91

In the table 2.5, racial and ethnic profiles of drivers arrested by police is showcased.

Table 2.5: Total Traffic arrests by Population Percentage

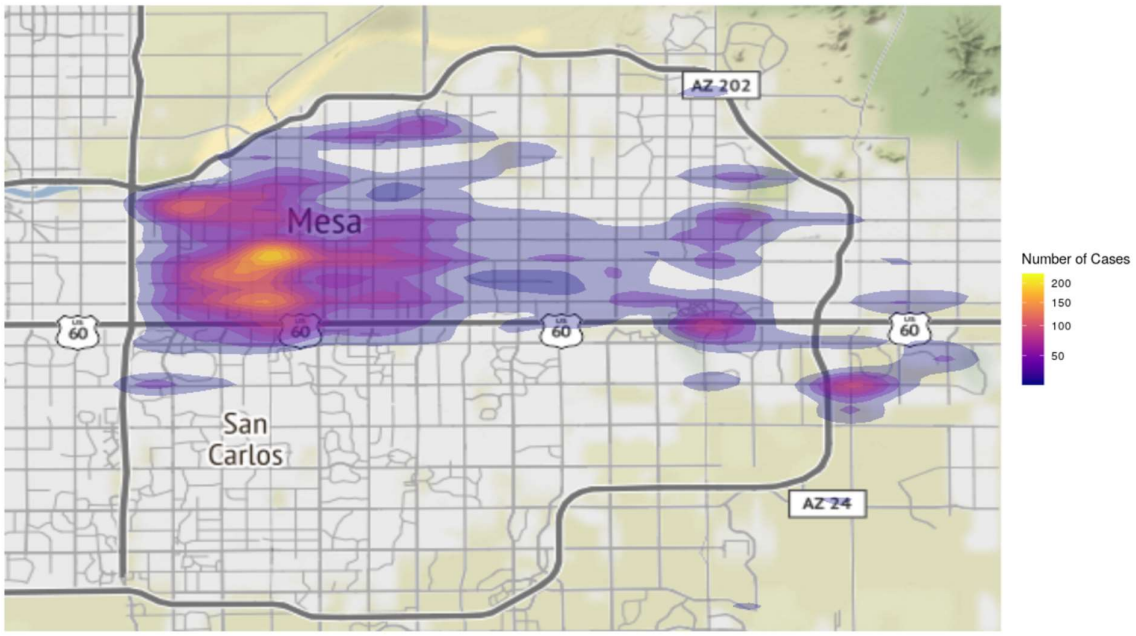
Driver's race	*Total Arrests*	*Population*	*Percentage Total Arrests*
Asian/pacific islander	5	45452	0.01
black	120	66911	0.18
Hispanic	302	540866	0.06
other	56	374025	0.01
white	391	1181965	0.03

Officers registered their view of the driver's race in one of five categories, with the percentage indicated in parentheses across the groups:

- White (65.05%)
- Hispanic (17.20%)
- Black (6.68%)
- Asian (1.3%)
- Other/Unknown race (9.72%)

It is important to remember that the differences in the percentages of ethnic groups that have been stopped across the city do not automatically indicate that police officers make race-based stop decisions. However, due to changes in the racial composition of residents and passengers, as well as differences in traffic flow patterns at these locations and possible differences in traffic violations, certain variations in the racial and ethnic background of stopped drivers in the city can be expected. There are significant differences in the ethnic background of the stopped drivers, as shown in Table 2.4.

Below is the density map for the hotspot for total number of cases around Mesa city.



Map 2.1 Density map for the number of cases in MESA city.

3 DATA PROCESSING:

Overview:

In my research, a machine learning method based on Spatial Analysis are introduced for optimum analysis and prediction of arrest violations at different locations. Multiple visualization techniques are used in this paper to better interpret the data, and further model the data for better results. This is achieved using R programming and the different packages it includes. Different phases in data processing are as follows.

3.1 Data Collection and Cleaning:

Usually, we collect this traffic stop data from government websites or police websites. But in this analysis data was obtained from Stanford Open Policing Project. Dataset for MESA city has been used in this research which has 96,621 records from 2014 to 2017. Data is loaded into R using Readr[10] package which is used to read rectangular data (like 'csv', 'tsv', and 'fwf') files (Wickham et al., 2018).

Data Cleaning is the process of transforming raw data into data that can be analysed. From the data obtained there were some raw data and some invalid values. So, to remove these NA values, I have used Tidy-verse package and some built-in functions.

3.2 Data Pre-processing:

In total there were 96,621 stops in 3 years, from that around 14,422 were removed for the reasons mentioned in Data section 2.2. From the remaining records there were around 6,000+ unique violations were recorded which committed by drivers around Mesa city.

For predicting the arrest violations at different locations, I have used machine learning models which I will be discussing in the methodology section. Here violation type acts as the response for my model. For my predictors, I have

included driver's characteristics (like sex, age, and race) as my predictors and I have also included location data(zipcode) in my model.

3.2.1 Violation Data:

As there were around 6000+ unique violations, I have categorised these violations based on Table 3.1. Using pattern matching and if else functions I divided these violations into 5 categories so that it will be easier to model the data when we have a smaller number of classes.

Table 3.1: Types of Traffic Tickets and their criteria

Violation Category	***Criteria***
Distracted Driving	Cell-phone usage/Drugs consumption/Liquor consumption or
Without proper documents or Vehicle defects	Drivers without driving license/expired license/Suspended license or Vehicles without registration number/expired insurance/Without proper windshield/Without child seat.
Leaving the Scene of an Accident	Leaving the scene of an accident without informing to police.
Reckless Driving	Reckless driving/ Driving on the Pedestrian walk/Not following lane rules/ Speeding
Running a Red Light or Stop Sign	Running over red light or stop sign

3.2.2 Location data:

From the data, we have location and their co-ordinate points. From the latitude and longitude values I extracted zip code data which gives neighbourhood

area codes around the city and included this data into the model as a location factor. I have used RevGeo package for generating zip-code from latitude and longitude values. RevGeo enables the use of the Photon geocoder for OpenStreetMap to reverse geocode coordinate pairs (Hudecheck, 2017)).

3.2.3 Unbalanced Data:

For modelling, I have used only arrest data for predicting arrest violation at different locations. There were around 900 stops which led to arrest and these data had a class imbalance. A difference in the frequencies of the observed groups can have a substantial negative impact on model fitting in classification problems. One method of solving such a class imbalance is to sub-sample the training data in a way that mitigates the problems. So, to balance this data I have used up-sampling method from caret package (Kuhn, 2020). Up-sampling will randomly sample (with replacement) the minority class to be the same size as the majority class.

3.3 Data Visualization:

Traffic stop data contain multiple entities. All these entities can be presented in terms of graphs providing representations of these stop records and relationships with time and driver characteristics etc. This section outlines the visualization techniques used to display the stop data.

3.3.1 Map Visualizations:

Given the geographic nature of the crime incidents, a Google-based interactive map was used for data visualization, where crime incidents are clustered according to their zip code detail. Ggmap and Ggplot2 packages has been used for showcasing the violation hotspot and many other maps (Kahle and Wickham).

3.3.2 Graph Visualizations:

In this paper, plots are made based on Ggplot2 and Plotly. It enables users to edit the plotting component at a high level that is compared to the basic R

graphs. Complex multilayer graphics can be produced easier by providing the powerful graphics models. The R graphing library of Plotly makes interactive graphs of the publication quality (Sievert, 2018).

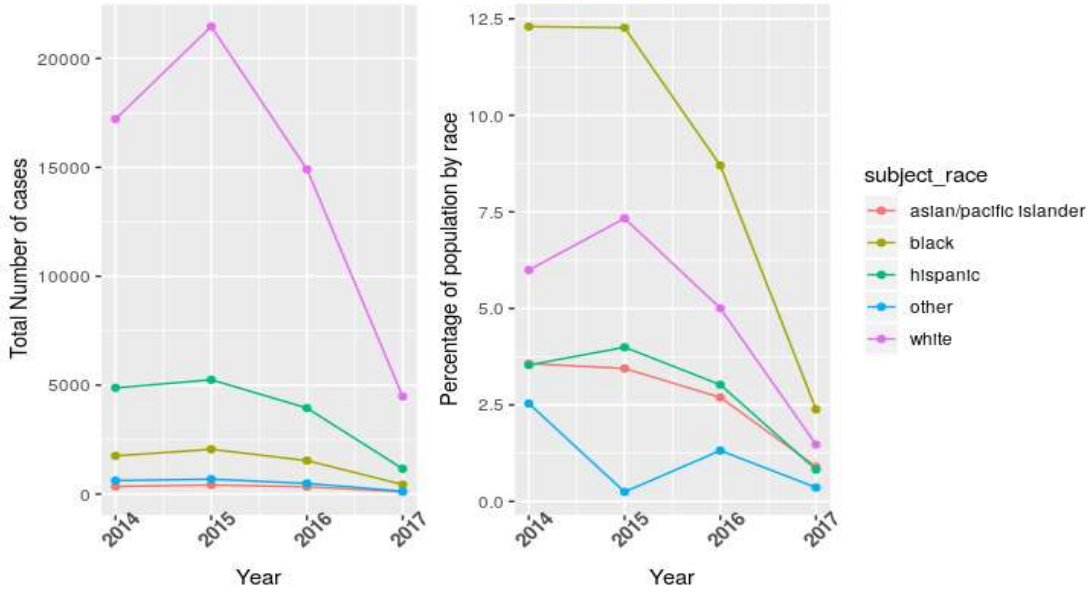
3.3.3 Shiny App:

In this paper, I have used a Shiny web application for showing results of a the best model which has been implemented. The application allows for user interaction and creates interactive map visualizations of the results. The application is developed using Shiny package in RStudio. Shiny is a web-based application framework for R which easily converts R (R Core Team, 2018) scripts into interactive GUI applications (Shiny - RStudio).

4 PRELIMINARY DATA ANALYSIS:

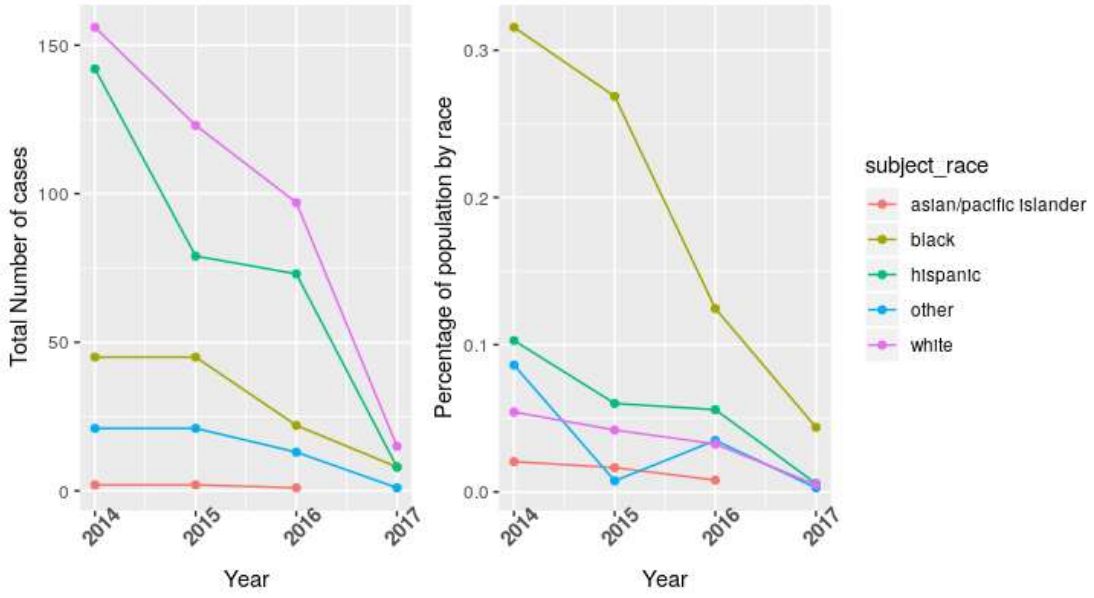
4.1 Trend over the years:

In this analysis, trend for being stopped at different locations are greater for White people when compared to other ethnic groups which can be noticed from Graph 4.1. But when compared to total population factor, I found out that percentage population for stopping Black people is more compared to White people.



Graph 4.1: Trend over the year for stopping people

Similarly, from graph 4.2 total number of arrests for white people is more compared to other races. But when we include population factor into the data, percentage population for arresting is greater for Black people. It is nearly 3 times greater than other races.

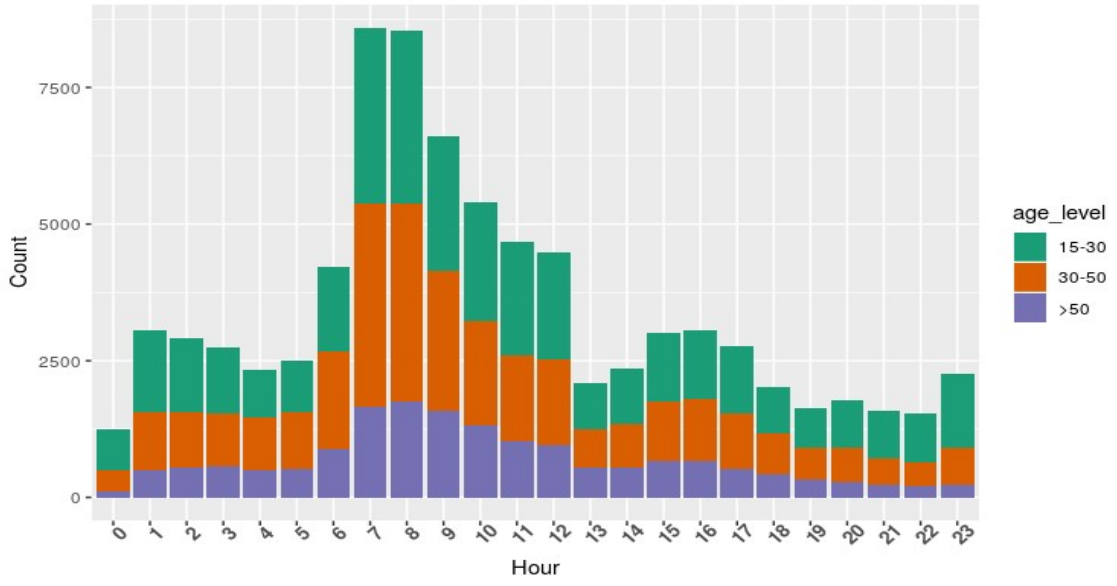


Graph 4.2: Trend over the year for arresting people

Bias towards black and Hispanic while stopping and arresting drivers exist (Pierson et al., 2020 and Goel et al., 2017). So, on comparing graphs, although number of stops and arrests for White people are greater compared to other races. This is because of the white population in the country. So, to overcome this population disproportionality, I extracted population data for Mesa city from US Census(U.S. Census Bureau QuickFacts: Mesa City, Arizona) and calculated population percentage by race. From the results, there is bias towards Black and Hispanic people in Mesa city as well. If police officers held Black and Hispanic people to the same standard as Whites, tens of thousands of searches of minorities might be avoided each year (Goel et al., 2017).

4.2 Total number of stops by hour:

On exploring, peak hour traffic in Mesa city is around 7-9am and 5-6pm (Traffic Patterns in Phoenix Arizona) was found. So, In the next analysis, have sorted data as total number of stops w.r.t. hour of the day. From the Graph 4.3, it can be shown that more number of cases are about 7-10am.

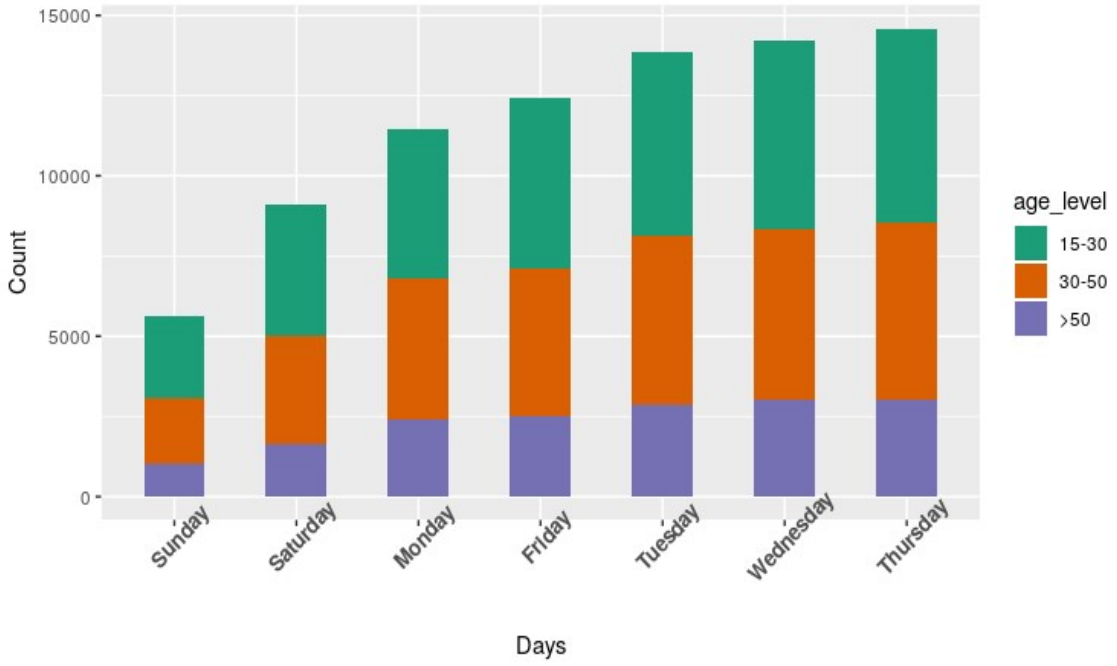


Graph 4.3: Total number of stops per hour for different age groups

From this, we can imply that more workers will head to their work early in the morning and will be in a rush to get to their workplace. They tend to do more violations and we can also notice that age group below 50 committed 80% of violations in peak hours. Peak hour traffic around 5-6pm cannot be explained from my analysis because the number of stops happened during these hours are less. One of the reasons for this is that number of violations committed by people will be less while heading back to their home from their work as they will have time to wait in traffic. So, the number of stops registered during these hours are less.

4.3 Total number of stops by Day:

In the next analysis, sorted the data for number of stops w.r.t. day of the week. From Graph 4.4, greater number of cases occur on weekdays compared to weekend.

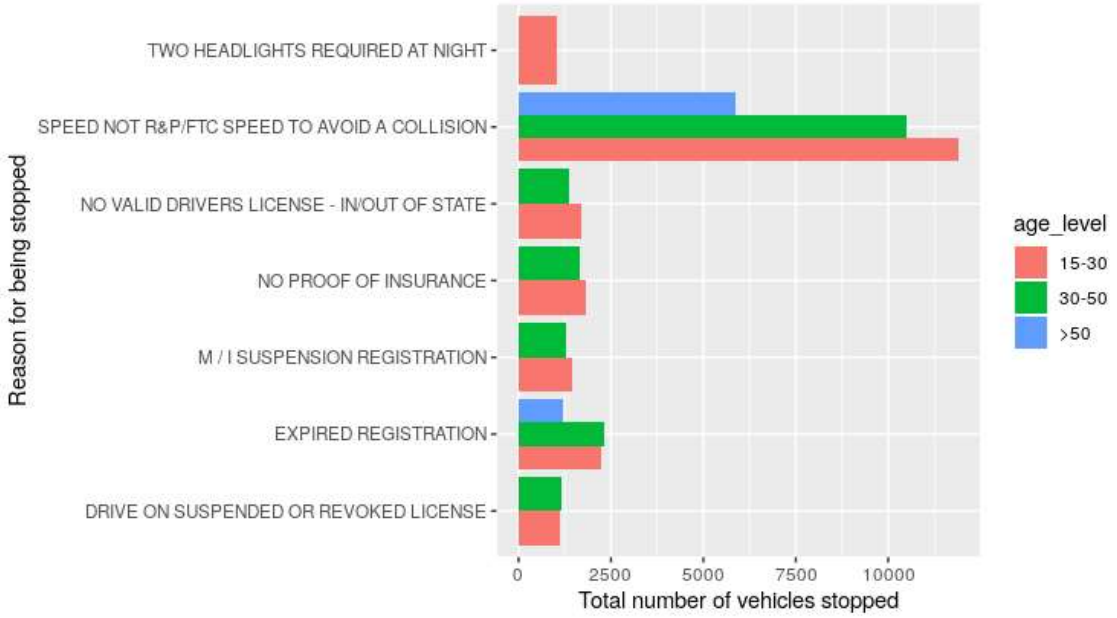


Graph 4.4: Total number of stops per day for different age groups.

From Graph 4.4 we can observe the same which matches the data present in "What Is the Busiest Travel Day" (Cristiano, Jason), that weekend traffic around the city is less compared to weekdays except on year's busiest travel days as the working days in a week is typically Mon-Friday.

4.4 Popular violation for getting stopped:

In this analysis, we show the sum of vehicles stopped for the most occured violation in MESA city.



Graph 4.5: Violation for being stopped vs Total number of stops

Analysing the traffic pattern around Mesa city, from website called access insurance, I found out that most common violation people commit in and around the city was Speeding violation(Traffic Patterns in Phoenix Arizona). From graph 4.5, popular violation in Mesa city is speeding which has 25,000+ cases which matches the data obtained from Traffic pattern. Here we note that under Speeding violation, more than 25,000 cars/trucks are stopped, in that more than 10,000 cars/trucks were driven by age category 15-30. Even though fines for this violation is more compared to other states (Az trial lawyer,2020), people tend to drive faster than the speed limit.

5 METHODOLOGY:

Overview:

In this paper , various machine learning methods based on Spatial Analysis are introduced for optimal analysis and prediction of arrest violations at different locations. In this research, violation column is divided into 5 categorical values like Distracted Driving, without proper documents or Vehicle defects, Leaving the Scene of an Accident, Reckless Driving, running a Red Light or Stop Sign. Here violation type acts as the response for my model. For my predictors, I have included driver's characteristics (like sex, age, and race) as my predictors and I have also included location data(zipcode) as a predictor in my model. The methods and various models are described as follows.

5.1 Multinomial Model:

The multinomial logistic regression model is an extension of the binomial logistic regression model. The independent variables can be either categorical (i.e., binary) or continuous (i.e., interval or ratio in scale). Binomial logistic regression model is further modified to categorize the dependent variable with nominal values of more than two levels using multinomial logistic regression. Most multivariate analyses include the basic assumptions of normality and continuous data, which include independent and/or dependent variables as stated above.

5.2 Decision Tree:

Decision tree is one of the well-known grading techniques. It builds tree structure classification and divides a dataset into smaller subsets. At the same time there is gradual creation of an related decision tree. The end result is a tree with nodes for decision and nodes for leaves. A node of decision has two branches, or more. Leaf node is a grouping, or a decision. The highest decision

node in a tree, which matches the best predictor named root node (Nasridinov et al., 2013 and Wang et al., 2012).

Decision Tree Algorithm works as follows

i. Place the best variable of the dataset at the root of the tree.

ii. Split training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an variable.

iii. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

For example, figure 5.1 shows how to determine when a consumer purchases a computer of a given kind.

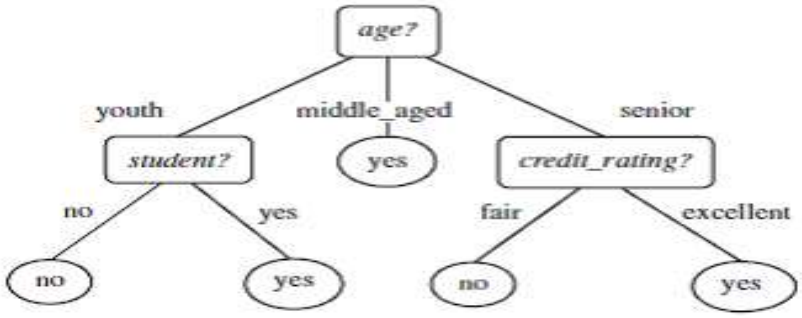


Figure 5.1 Decision tree Example (ZhengTianyu, 2013)

5.3 Random Forest:

Random forest builds and merges multiple decision trees to obtain a more accurate and stable predictions. In this model several decision trees are built using subsets drawn from the training set. The splitting of each tree node is not based on the best splits of all characteristics, but rather on the best split among a random number of features. The tree bias increases because of randomness, but averaging also helps to decrease variance, which is why this model often achieves better results. Figure 5.2 represents an example for how random forest works. This image was obtained from Wikipedia (Random Forest).

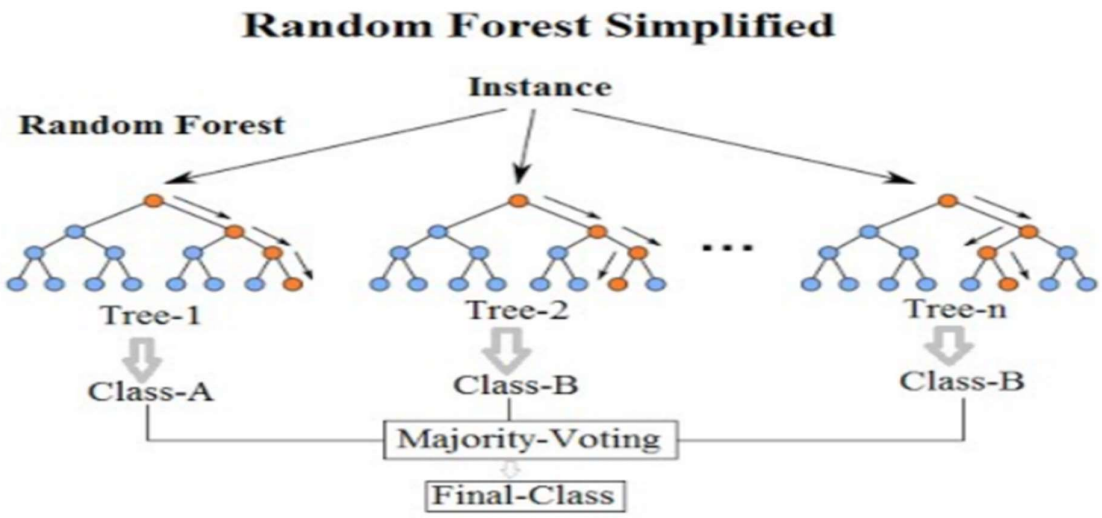


Figure 5.2: Random Forest workflow (Random Forest)

5.4 Neural Networks:

The Artificial Neural Networks (ANN) were developed as generalizations of biological nervous system mathematical models. Neural network 's main computing components are called artificial neurons, or simply neurons or nodes. Artificial neurons initially aimed at modeling biological neurons. Nowadays we are only interested in finding models that produce the best results for the tasks of Machine Learning. An ANN is composed of a series of interconnected processing components, also known as neurons or nodes. It can be described as a directional graph in which each node executes a form transfer function.

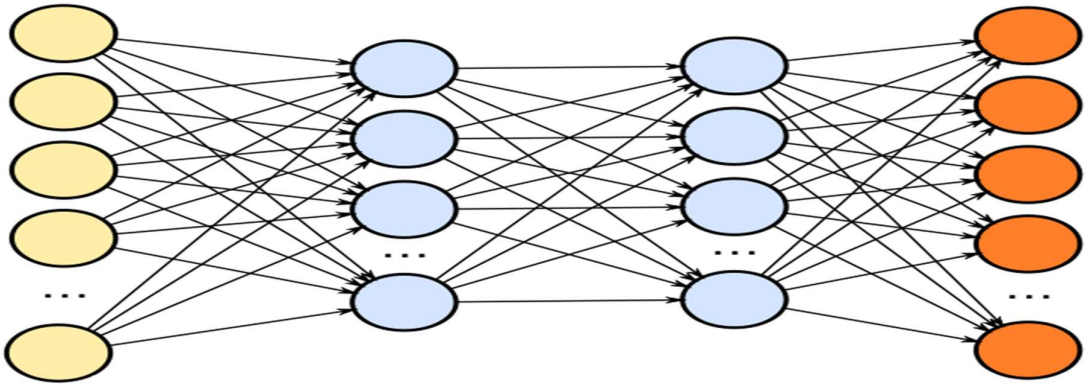


Figure 5.3: Neural network Architecture (Tabacof et al., 2016)

5.5 K-Nearest Neighbour:

The theory behind the k nearest neighbour classification is to apply the Bayes rule to define and identify the test observation with the highest probability. This corresponds to taking a majority vote of the data point's closest K neighbours and choosing that class as the response that has the highest number of votes. The knn algorithm in R requires quantitative predictors. It is possible to construct a version using categorical predictors. The response for knn must be categorical, there is no limitation on the number of categories

Figure 5.4 illustrates how it works in a clear way. We need to find the neighbours and contact out which animal this could be. If k=1, the only neighbor is a cat, and assumes the piece of the puzzle should also be a cat. When k=4, one chicken and three cats are in the closest neighbourhood. Also, it's fair to say it would be a cat in this situation.

Figure 5.4: K-nearest neighbour Example (Python Machine Learning Tutorial)

5.6 Performance metrics:

Testing the machine learning models, there are certain terminologies that one needs to understand (Billa, 2019).

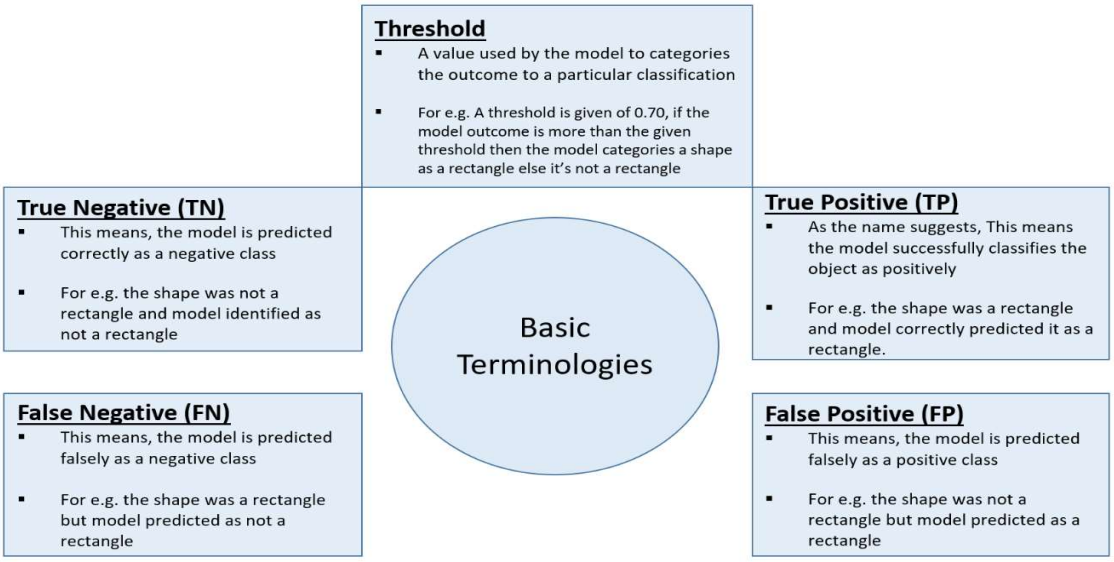


Figure 5.5: Terminologies for Confusion Matrix

Classification Accuracy: This is the common way of evaluating the machine learning models. It is a ratio between the positive predictions vs the total number of predictions. If the value is high, then the model has a high prediction accuracy. Below are the formulas for accuracy.

Equation 1: Accuracy for the model

$$Accuracy = \frac{Total\ Positive\ Prediction}{Total\ Number\ of\ Predictions}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

It is also shown that accuracy alone is not a reasonable way of testing the model.

Confusion Matrix: A confusion matrix is a table often used to define a classification model 's output on a collection of test data for which the true values are identified. Below is a list of values for a classifier which are sometimes derived from a confusion matrix:

- True Positive Rate: When it is actually yes, how often does it predict yes? also known as "Sensitivity" or "Recall"
- Precision: When it predicts yes, how often is it correct?
- F1 Score: This is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than the Accuracy Metric.

Equation 2: Formulas for Precision, Recall and F1-score

$$Precision = \frac{TP}{TP+}$$

$$Recall\ = \frac{TP}{TP+FN}$$

$$F1\ Score = \ 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

6 RESULTS:

Below are some interesting results from the analysis.

6.1 Multinomial model

This model is trained for full dataset as there were only 900 records for only arrest cases and performed K-fold cross validation. Data contains stops from January 1, 2014 to Dec 31, 2017. The model would predict the arrest violation category using the driver's characteristics (age, race, and sex) and the district (Zip-code) where it happened as predictors.

Table 6.1 represents confusion matrix for Multinomial Model which shows the data about predicted classes from the model.

Table 6.1: Confusion Matrix for Multinomial Model

Violation Type	***distracted_driving***	***leaving_scene_of_an_accident_informing***	***reckless_driving***	***running_red_light_or_without_headlight***	***without_proper_documents_or_vehicle_defects***
distracted_driving	26	2	34	25	142
leaving_scene_of_an_accident_informing	7	11	30	21	71
reckless_driving	5	0	26	13	86
running_red_light_or_without_headlight	10	1	30	47	85
without_proper_documents_or_vehicle_defects	10	2	28	18	140

6.2 Decision Tree:

In this section, we see how decision tree fits our data. Selecting Predictors is the main step to build a decision tree. I have used predictors such as driver's characteristics (age, race, and sex) and the district (Zip-code). These attributes indicate various Mesa stops from year 2014-2017. I classified this violation data into five classes which acts as response for my model. Using this method, we

have obtained a classification tree shown in Fig. 6.1. In this tree, nodes belong to the violation data attributes. For example, if the driver's age is greater than 25.5 and driver's race is Hispanic, this tree gives percentage of each violation which gets him/her arrested. In that only one violation has 90 percent probability and rest all violations has less than 10 percent probability.

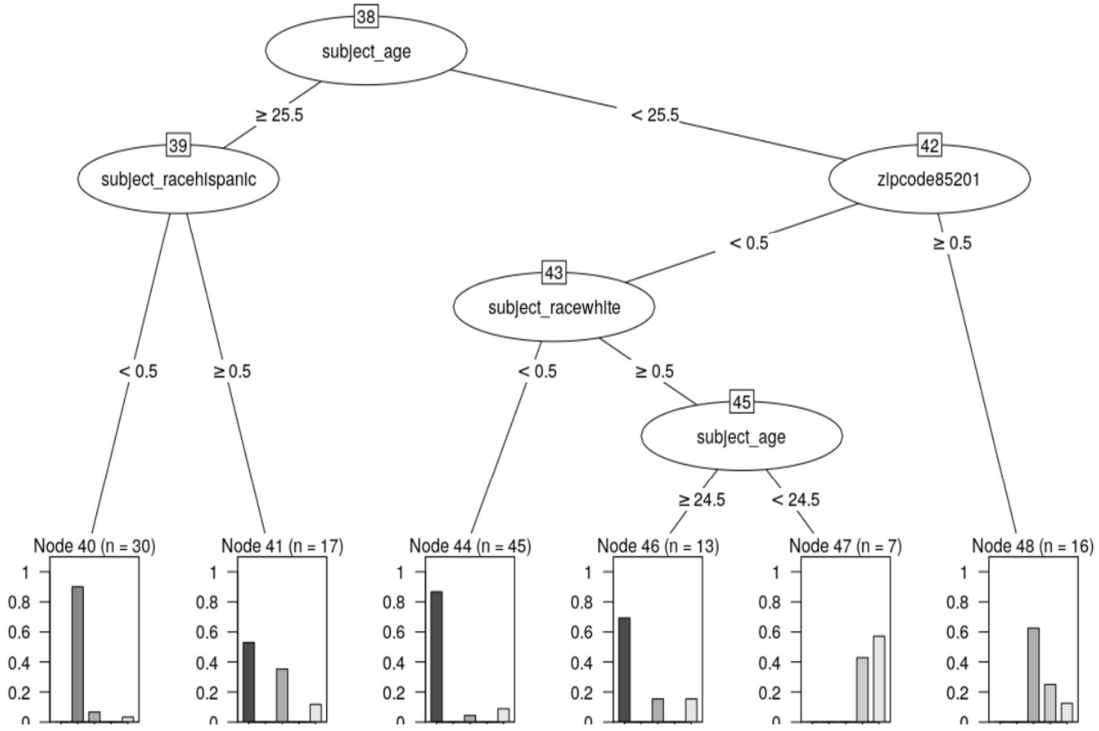


Figure 6.1: Subset of Decision Tree

Table 6.2 represents confusion matrix for Decision Tree which shows the data about predicted classes from the model

Table 6.2: Confusion matrix for Decision tree

Violation Type	***distracted_driving***	***leaving_scene_of_an_accident_informing***	***reckless_driving***	***running_red_light_or_without_headlight***	***without_proper_documents_or_vehicle_defects***
distracted_driving	52	0	17	9	76
leaving_scene_of_an_accident_informing	1	16	5	4	10
reckless_driving	1	0	69	9	68

running_red_light_or_without_headlight	2	0	26	87	101
without_proper_documents_or_vehicle_defects	2	0	31	15	269

6.3 Random Forest:

In this model, model training is done in the same way as for individual decision trees. It is the automatic selection of variables in data that are most relevant to the modelling problem. The importance of each feature variable in a training subset refers to the portion of the gain ratio of the variable compared with the total feature variable. The value of all feature variables is sorted in descending order and the top variable values are selected. Here at each split 27 variables has been used and 500 number of trees were trained for better accuracy. Out of Bag estimate of error rate is 15.57%. Confusion matrix is shown in Table 6.3.

Table 6.3: Confusion Matrix for Random forest

Violation Type	***distracted_driving***	***leaving_scene_of_an_accident_informing***	***reckless_driving***	***running_red_light_or_without_headlight***	***without_proper_documents_or_vehicle_defects***
distracted_driving	58	0	11	6	31
leaving_scene_of_an_accident_informing	0	16	1	3	6
reckless_driving	0	0	128	5	37
running_red_light_or_without_headlight	0	0	7	108	40
without_proper_documents_or_vehicle_defects	0	0	1	2	410

6.4 Neural Net:

This neural network model was designed and studied to classify the violation types based on categorical input data containing spatial information on the violation. To develop a neural network-based classifier, a fully connected feed-forward model was developed and its hyperparameters were optimized to avoid underfitting and increase the accuracy. As a result, the accuracy of this model was 43%. A 27-5-5 network with 170 weights model was fitted for my data. Table 6.4 represents confusion matrix for neural net model.

Table 6.4: Confusion Matrix for Neural net

Violation Type	***distracted_driving***	***leaving_scene_of_an_accident_informing***	***reckless_driving***	***running_red_light_or_without_headlight***	***without_proper_documents_or_vehicle_defects***
distracted_driving	30	0	13	11	53
leaving_scene_of_an_accident_informing	1	15	3	4	17
reckless_driving	5	1	70	22	147
running_red_light_or_without_headlight	4	0	21	66	112
without_proper_documents_or_vehicle_defects	18	0	41	21	195

6.5 K-Nearest Neighbour:

Violation prediction by using the k-nearest neighbour. This is one of the simplest models. Using predictors like driver's characteristics (age, race, and sex) and the district (Zip-code) where it happened is found to be 47% accuracy with k value as 5. Table 6.5 represents confusion matrix for K-nn model.

Table 6.5: Confusion matrix for K nearest neighbour:

Violation Type	distracted_driving	leaving_scene_of_an_accident_informing	reckless_driving	running_red_light_or_without_headlight	without_proper_documents_or_vehicle_defects
distracted_driving	58	0	11	6	31
leaving_scene_of_an_accident_informing	0	16	1	3	6
reckless_driving	0	0	128	5	37
running_red_light_or_without_headlight	0	0	7	108	40
without_proper_documents_or_vehicle_defects	0	0	1	2	410

6.6 Performance Analysis:

Performance metrics like F1 score, accuracy, recall, precision of each model is calculated using Confusion Matrix function. Table 6.6 provides these data from the models that I have fitted.

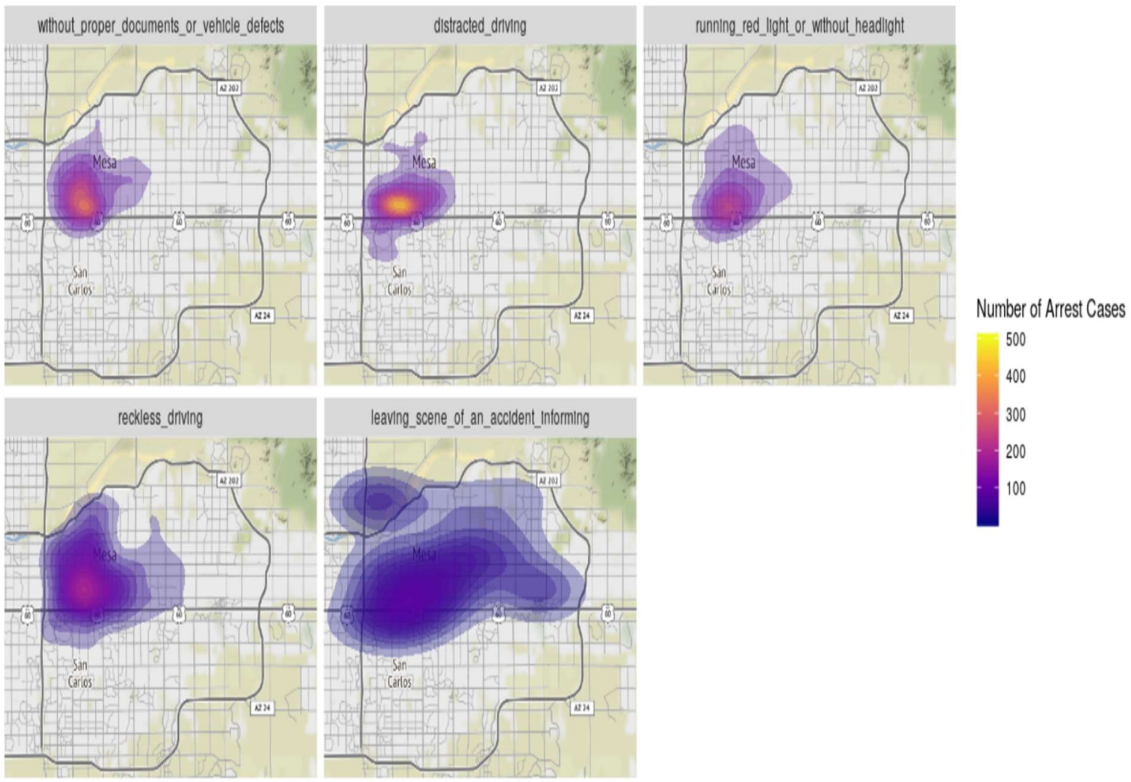
Table 6.6: Performance metrics table

Model Type	Accuracy	Recall	Precision	F1
Multinomial	**34.02%**	**41.14%**	**29.22%**	**34.17%**
Decision Tree	**56.67%**	**71.55%**	**50.06%**	**58.91%**
Random Forest	**83.10%**	**90.38%**	**72.37%**	**80.38%**
Neural Net	**43.22%**	**56.64%**	**39.51%**	**46.55%**
K-nearest neighbour	**47.24%**	**73.44%**	**48.08%**	**58.12%**

From comparing all 5 models, random forest model has the best Accuracy and F1 value.

6.7 Map Visualization of Predicted violation:

The arrest violations can be graphically represented using a density map which indicates number of arrests, brighter colour to indicate less number of arrests and lighter colours to indicate more number of arrests. Map 6.1 is an example of a heat map for Mesa city. In this map, hotspot for each arrest violation type can be found. This heat map for each arrest violation type is predicted from Random Forest model.



Map 6.1: Density map for different arrest violation classification (Violation hotspot)

6.8 Shiny App for Model Output:

From Random forest model, probabilities of each arrest violations are predicted and created a shiny app for visualising these results, based on violation type, driver's race, and driver's sex. Here we can check gender-based or racial based arrests made in Mesa city. In this app, it allows the user to check percentage of each arrest violation type at different locations based on driver's race, sex the

violation type one has committed, Figures 6.2 show screenshot from the Shiny app. Notice the option to analyse various violation types, different race, and gender selection are included in the app. The application represents a map of the crime category which the user selects from a drop-down list, a column of the data of selected variable.

This application shows a map of the Mesa city where each point represents each location and percentage for each arrest violation type is displayed on each point. Link for this shiny can be found here.

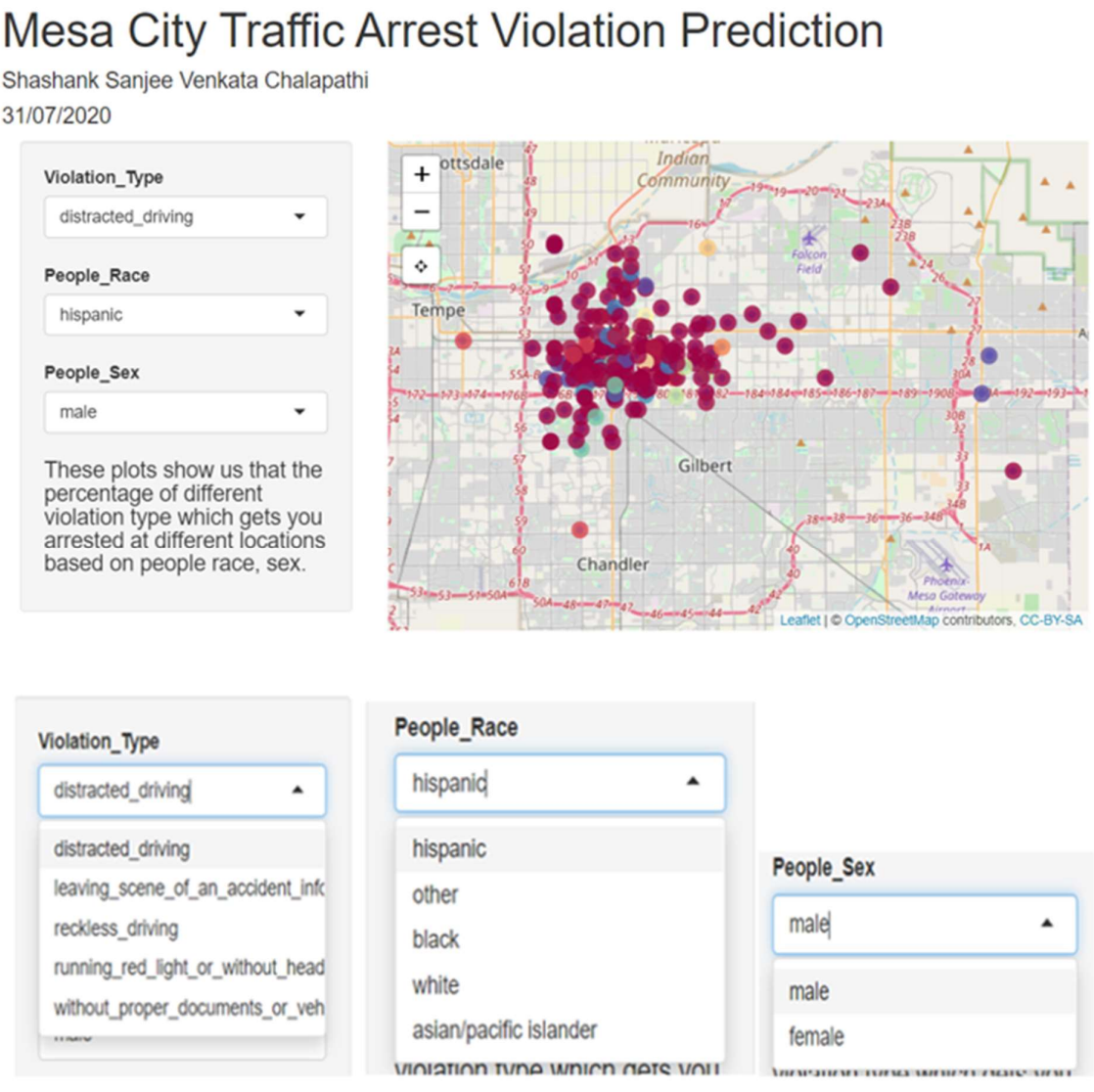


Figure 6.2: Screenshot taken from the shiny app for model output

7 DISCUSSION:

Internal analysis which compares the racial bias of traffic stops and arrests across different locations. So, to further analyse this, my report focused on whether racial/ethnic disparities are evident in post-stop outcomes i.e. arrests based on location, race and sex.

Arrest Violation prediction by using various machine learning methods. These methods provide an overview of the large amount of violation data and promote the storage, scanning and retrieval at various locations of the required violation information . Arrest Violation based on different locations can help the police officers to take proper measures against that violation. For example, zip-code 85210 area has the highest number of arrests for violation "without driving license or without proper vehicle documents" around Mesa city. So police officials can take proper precautions and rules in that area for this particular violation.

Using the intelligent traffic accident detection system, supervised machine learning algorithms such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Random Forests (RF) are used on traffic data to build a model to identify accident cases from normal cases. The intelligent traffic accident detection system which uses simulated data from adhoc vehicle networks (VANETs) based on vehicle speeds and coordinates and then sends driver traffic alerts (NejdetDogru, 2018). This also demonstrates how to manipulate machine learning methods to predict highway accidents in ITS. The efficiency of the RF algorithm was found to be superior to the ANN and SVM algorithms in terms of its precision for violation prediction on highways.

From "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis" (Weiweilin, 2017), that deals with the imbalanced distribution of business data, the lack of user functionality and many other factors, the direct use of big data strategies on practical business data appears to deviate from the

business objectives. It is difficult to model data from the insurance industry through classic algorithms like Logistic Regression and SVM etc. The result of the experiment shows that SVM and other classification algorithms were outperformed by the ensemble random forest algorithm in both performance and accuracy within the imbalanced results. So to deal with imbalance data random forest performs better than other models.

Among 5 models which I have fitted for the data. I evaluated each of the 5 constructed models separately. Multinomial model is our first model that was used to extract arrest violations at different location which has less accuracy and F1 score. Next on comparing accuracy and f1 score of other models from table 5.6. We can conclude that Random forest model has the highest accuracy and f1 score. Upon that Random Forest model possess low bias and moderate variance. Also, it handles unbalanced data very well. Even though it takes more space for model to fit, performance wise it gives accurate results. So, from random forest model predictions at each location is collected used for this analysis.

Using the shiny app, one can avoid taking those routes based on their race and sex. Using this app police officials can reduce arrests by replacing officers near crime hotspot to reduce racial bias while arresting/stopping. Like the system developed by IBM "Crime Prediction and Prevention", which helps in reducing crime, saving time and optimize resources by using data driven insights to help know what's coming and taking advantage of structured and unstructured data sources including incident reports, surveillance, sensor, and social media content (Crime Prediction and Prevention, IBM).

So based on this prediction model, police officers can increase the speed detection devices in and around the City. Also increasing fines so that people might stop committing the violations. In addition to that, increasing police patrolling on respective areas where the more stops are being made can reduce the number of violations in crime hotspots around the city.

8 CONCLUSION:

In this report, I generated many graphs and found interesting results that showed the basis for understanding the data sets of Mesa city traffic stops . Racial inequalities in traffic stop outcomes were noticed even after including other factors, such as demographic factor believed to affect police decision making during traffic stops, were taken into account. The importance of these findings which lead to stops resulting in arrests. These statistics cannot directly determine the reasons for the existing racial differences identified for traffic stop outcomes in the city. The models can only calculate the influence of the variables on stop outcomes. Therefore, no statistical methods can assess the intent of the officers or individual biases.

Next, we applied machine learning models on arrest data to find frequent arrest violation patterns in the city at different locations, to help predicting future arrest violations in a specific location. We achieved 83% of prediction accuracy in Mesa city for random forest. We aimed to further understand our models results and to predict arrest violation hotspot around the city. Finally, we provided shiny app for analysing model results so that police officers can increase the speed detection devices in and around the City.

9 FURTHER ANALYSIS:

We plan to apply more classification models to improve the accuracy of crime prediction and the overall performance . It is also a helpful extension for our study to consider the population information area-wise to see if there are actual racial bias in that particular areas. Furthermore, we want to study these same methods for other stop datasets from new cities along.

10 REFERENCES:

1. "Crime Prediction and Prevention." IBM, www.ibm.com/industries/government/public-safety/crime-prediction-prevention.
2. "Python Machine Learning Tutorial." Machine Learning with Python: k-Nearest Neighbor Classifier in Python, www.python-course.eu/k_nearest_neighbor_classifier.php.
3. "Random Forest." Wikipedia, Wikimedia Foundation, 11 July 2020, en.wikipedia.org/wiki/Random_forest.
4. "Types of Traffic Tickets." Findlaw, https://traffic.findlaw.com/traffic-tickets/types-of-tickets.html
5. "U.S. Census Bureau QuickFacts: Mesa City, Arizona." Census Bureau QuickFacts, www.census.gov/quickfacts/mesacityarizona.
6. Applied Statistics, Social Science, and Humanities, New York University, New York, NY, USA.
7. Az trial lawyer. "Speeding in Arizona May Cost You a Fortune: The Law Office of Karl A. Mueller, PLC." Aztriallawyer, 3 July 2020, www.aztriallawyer.com/2018/07/speeding-in-arizona-may-cost-you-a-fortune/
8. Billa, Mukund. "Testers Guide for Testing Machine Learning Models." Medium, Analytics Vidhya, 18 Oct. 2019, medium.com/analytics-vidhya/testers-guide-for-testing-machine-learning-models-e7e5cea81264.
9. Communication, Stanford University, Stanford, CA, USA.
10. Computer Science, Stanford University, Stanford, CA, USA.
11. Cristiano, Jason. "What Is the Busiest Travel Day?" traveltips.usatoday.com, https://traveltips.usatoday.com/busiest-travel-day-108182.html. 19 July 2020.
12. Davis, Elizabeth, et al. "Bureau of Justice Statistics Home Page." Bureau of Justice Statistics (BJS), www.bjs.gov/index.cfm?ty=dcdetail&iid=251

13. Goel, S., & Phillips, C. (2017, June 19). Police Data Suggests Black and Hispanic Drivers Are Searched More Often Than Whites. Retrieved June 29, 2020, from https://slate.com/technology/2017/06/statistical-analysis-of-data-from-20-states-suggests-evidence-of-racially-biased-policing.html
14. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
15. Hudecheck, Michael (2017). revgeo: Reverse Geocoding with the Photon Geocoder for OpenStreetMap, Google Maps, and Bing. R package version 0.15. https://CRAN.R-project.org/package=revgeo
16. Kahle and Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
17. Kuhn, Max (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
18. Management Science & Engineering, Stanford University, Stanford, CA, USA. [Accessed 24 July 2020]
19. Nasridinov, Aziz & Ihm, Sun-Young & Park, Y.-H. (2013). A Decision Tree-Based Classification Model for Crime Prediction. 253. 531-538. 10.1007/978-94-007-6996-0-56.
20. NejdetDogru, “Traffic Accident Detection Using Random Forest Classifier”,978-1-5386-2659-7/18/$31.00 ©2018 IEEE.5.
21. Pedro Tabacof, Julia Tavares, and Eduardo Valle. (2016). Adversarial Images for Variational Autoencoders.
22. Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., . . . Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. Nature Human Behaviour. DOI:10.1038/s41562-020-0858-1
23. Shiny - RStudio. URL: https://www.rstudio.com/products/shiny/

24. Sievert, Carson (2018) plotly for R. https://plotly-r.com
25. Starr, Sonja B. (2016) "Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police," University of Chicago Legal Forum: Vol. 2016, Article 12. Available at: http://chicagounbound.uchicago.edu/uclf/vol2016/iss1/12
26. Traffic Patterns in Phoenix Arizona, 12 July 2019, www.access-insurance.com/about-us/blog/traffic-patterns-phoenix-arizona. [Accessed 24 July 2020]
27. Traffic Patterns in Phoenix Arizona, www.access-insurance.com/about-us/blog/traffic-patterns-phoenix-arizona. [Accessed 1 August 2020]
28. Uchida, C., Swatt, M., Gamero, D., Lopez, J. Salazar, E., King, E., … Michael, D. (2012). Los Angeles California smart policing initiative (pp. 1–12). Washington, DC: Bureau of Justice Assistance, US Department of Justice. [Accessed 24 July 2020]
29. Wang B, Dong H, Boedihardjo AP, Lu CT, Yu H, Chen IR, Dai J (2012) An integrated framework for spatio-temporal-textual search and mining. In: Proceeding of the 20th SIGSPATIAL international conference on advances in geographical information systems, pp 570–573 [Accessed 6 August 2020]
30. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
31. ZhengTianyu, 2013. Decision Tree And Entropy Algorithm. ZhengTianyu's Blog. Available at: zhengtianyu.wordpress.com/2013/12/13/decision-trees-and-entropy-algorithm/ [Accessed 6 August 2020].

11 APPENDIX:

11.1 Shiny Application Code:

```
##########################################
##
##  Mesa City Traffic Violation Prediction
##
##########################################

#########################
# Required Libraries
#########################
library(tidyverse)
library(ggmap)
library(shiny)
library(ggplot2)
library(leaflet)
library(readr)
library(rsconnect)

rsconnect::setAccountInfo(name = 'shashanksv',
                          token = 'D35842D957D06925687CD85C46AFE987',
                          secret =
'WKfa1YbaLaF88q/cyUx4P8kid6lquJWuSzOGXlNZ')

shiny_data <- read_csv("shiny_data.csv")

vio <- unique(shiny_data$violation_type)
age <- unique(shiny_data$subject_age)
race <- unique(shiny_data$subject_race)
sex <- unique(shiny_data$subject_sex)

###### UI function ######
ui <- fluidPage(sidebarLayout(
  sidebarPanel(
    selectInput(
      "Violation_Type",
      label = "Violation_Type",
      choices = vio,
      selected = vio[1]
    ),
    selectInput(
      "People_Race",
      label = "People_Race",
      choices = race,
      selected = race[1]
    ),
    selectInput(
      "People_Sex",
      label = "People_Sex",
      choices = sex,
      selected = sex[1]
    ),
    h4(
      "These plots show us that the percentage of different violation type
which gets you arrested at different locations based on people race, sex."
    )
  ),
  mainPanel(leafletOutput("bbmap"))
))
```

```
###### Server Function ######
server <- function(input, output, session) {
  shiny_data$popup <-
    paste(
      "<br>",
      "<b>Violation type: </b>",
      shiny_data$violation_type,
      "<br>",
      "<b>Percentage: </b>",
      shiny_data$prob * 100
    )
  pal <- colorNumeric(palette = "Spectral",
                      domain = shiny_data$prob)

  output$bbmap <- renderLeaflet({
    shiny_data1 <-
      shiny_data %>% filter(
        violation_type == input$Violation_Type,
        subject_race == input$People_Race,
        subject_sex == input$People_Sex
      )

    leaflet(shiny_data1) %>%
      addCircles(lng = ~ lng, lat = ~ lat) %>%
      addTiles() %>%
      addCircleMarkers(
        data = shiny_data1,
        lat =  ~ lat,
        lng = ~ lng,
        radius = 7,
        popup = ~ as.character(popup),
        color = ~ pal(shiny_data1$prob),
        stroke = FALSE,
        fillOpacity = 0.8
      ) %>%
      addEasyButton(easyButton(
        icon = "fa-crosshairs",
        title = "ME",
        onClick = JS("function(btn, map){ map.locate({setView: true}); }")
      ))
  })

}

###### Calling Shiny Application ######
shinyApp(ui, server,
         options = list(height = 600))
```