Large-Scale Analysis of Citation Bias in San Diego (May 2019)

Michael Reed, Salman Alsalman, and supervised by Dr. Xiaobai Liu

Computer Science Department
San Diego State University

INTRODUCTION

The police are a mainstay in broader American society and culture where many see them as domestic protectors and arbitrators of the American citizenry [1]. However, recent events such as the Ferguson riots and the shooting of a child named Tamir Rice has brought the question of the legitimacy of police to the public discourse. Do police, who are supposed to be neutral protectors, suffer from personal bias in their interactions with the denizens of the towns and cities they patrol? The Department of Justice investigation into the Ferguson Police Department showed that the police were one of the main pipelines for generating municipal revenue for a city that has been laden with financial difficulties. The main tools of collecting revenue was through issuing citations and tickets [2]. The most common means for citizens to interact with the police and thus receive fines is through traffic stops. These interactions provide us with data to assess the behavior of police [3]. While large scale analysis of data pertaining to police is difficult primarily due to availability, the Stanford Open Policing Project has collected data on traffic stops conducted throughout the country [4]. In our project we aim to develop a prediction system that will assess whether a citizen will be issued a citation based on other factors about the driver including age, sex and time of the stop.

Index Terms— Decision Making, Citations Issued, Classification, Prediction, Decision Tree, Neural Network, Random Forest.

Task description

The main goal of our research is to analyze, study, and find if there is a relationship between deciding to give a citation to a driver, and their race. Also, to build machine learning classifiers and predictors to assess the study presented. The scope of our study is for the city of San Diego. To achieve this goal, the following steps will be done.

1. Acquiring the data from its source and understand it.
2. Explore the nature of the data and their values' datatypes.
3. Preprocess the data and prepare it for processing.
4. Create and building machine learning models for our study.
5. Evaluate the performance of our predictors' models.

EXPERIMENT

Dataset Description

The dataset we are using, represents a police traffic stops in San Diego. This dataset is a stripped-down of a larger dataset that covers the whole country. The original dataset has around 93 million records. It includes all the stops for six years period, from 2011 to 2017. There are four major cities from California in the original dataset: Bakersfield, San Diego, San Francisco, and San Jose. The total number of entries for California is 1,111,573 stops. San Diego itself has 390,867 stops, which is 35.2% of the whole California entries.
In our study, we will use the dataset that has been released for the city of San Diego only. We got it from Stanford Open Policing Data Website [4].

THE STANFORD OPEN POLICING PROJECT:
https://openpolicing.stanford.edu/data/

The dataset consists of 19 columns and 390,867 rows. Table 1 shows brief description of these columns.

COLUMN NAME	COLUMN DATATYPE	DESCRIPTION
RAW_ROW_NUMBER	Integer	A sequencing for the rows
DATE	String	Police stop date
TIME	String	Police stop time
SERVICE_AREA	Integer	Stop location
SUBJECT_AGE	Real Number	Driver Age
SUBJECT_RACE	Categorical	Driver Race
SUBJECT_SEX	Categorical	Driver Sex
TYPE	Categorical	Vehicle Type
ARREST_MADE	Categorical	Arrest decision made or not
CITATION_ISSUED	Categorical	Citation decision issued or not
WARNING_ISSUED	Categorical	Warning decision issued or not
OUTCOME	Categorical	The decision made after the stop
CONTRABAND_FOUND	Categorical	Contraband found or not
SEARCH_CONDUCTED	Categorical	Was a search been conducted
SEARCH_PERSON	Categorical	Was a personal search happened
SEARCH_VEHICLE	Categorical	Did the vehicle been searched
SEARCH_BASIS	Categorical	Search basis
REASON_FOR_SEARCH	Categorical	The reasons for a search
REASON_FOR_STOP	Categorical	The main reason of the stop

The goal of our study is to find out if giving a citation decision or not, might be related to the driver's ethnicity. There are two features in our dataset, that logically affect this study; 'subject_race' and 'citation_issued'. 'Subject race' feature is a multi-valued categorical column. However, 'Citation issued' feature is a truth-valued binary column, where the values might be true or false. Thus, we decided to choose 'citation_issued' feature as the selector column for our predictors.

Data Exploring and Preprocessing

Data Preprocessing

It is important to know that each dataset may contain values, columns, rows, and samples that might even affect exploring the nature of the dataset. That is why the phase of exploring the data could be done in two phases: pre-explore and explore the data. Between these two phases we need to do pre-processing and cleaning of the dataset. One of the big challenges in dealing with big datasets, is having null values or unknown values. These values might lead to wrong prediction and classification tasks.

The dataset we have has 390,999 entries and 19 columns. It has many null values within it. These nulls will affect our goal of making decision predictors based on our dataset.

We counted the number of null values in our dataset and found that it has 2,077,360 null entries. This huge number represent 27.96 % of the whole dataset entries. We did column-wise analysis to detect which column has high null values; the following table shows the results.

Column Name	*Null Values Rate*
raw_row_number	0.000000
date	0.033760
time	0.321228
service_area	0.000000
subject_age	3.233768
subject_race	0.357546
subject_sex	0.206139
type	0.000000
arrest_made	8.957056
citation_issued	8.366262
warning_issued	8.366262
outcome	10.242226
contraband_found	97.144750
search_conducted	9.487492
search_person	96.281320
search_vehicle	96.281320
search_basis	95.696664
reason_for_search	96.251653
reason_for_stop	0.068031

Thus, in order to clean our dataset from the null values, we decided to do the following steps:

1. Remove columns with high null values rate.
2. Remove columns that have one value entry such as ‘type’ column that has the value ‘vehicular’ in all its rows.
3. Remove unnecessary columns such as ‘raw_row_number’ because it is only representing the raw sequencing of the dataset's entries.
4. Remove rows that have null values.

By doing the first three steps, the number of columns to be processed dropped from 19 to 12, and the null values rate decreased from 27.96% to 4.14%. Then, after removing the rows with nulls, the null values rate reached 0.0%.

Obviously, we have cleared our dataset of all the null values. However, more preprocessing steps are necessary in order to get the best version of the dataset that will work the best for our study and predictors. The following steps were also performed in the preprocessing phase:

1. Remove rows with 'Unknown' value in service area entries. These values are unknown, mostly, due to human errors.
2. Reformat Entries Values. For our study and to ease the complexity of the data, we decided to reformat the values of the time and date of the stops.

The following table shows part of the dataset to be used in our study.

ject_sex	arrest_made	citation_issued	warning_issued	outcome	search_conducted	reason_for_stop
male	False	True	False	citation	False	Moving Violation
male	False	False	True	warning	False	Moving Violation
male	False	False	True	warning	False	Moving Violation
male	False	True	False	citation	False	Moving Violation
male	False	True	False	citation	False	Equipment Violation

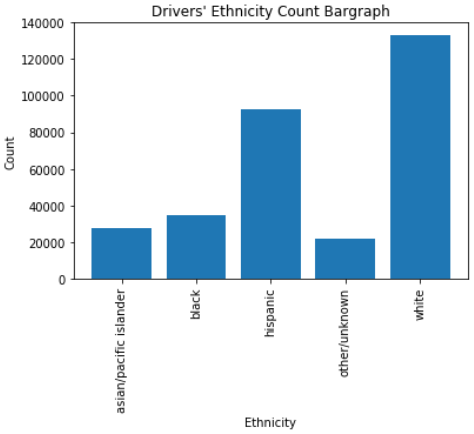
Data Transformation

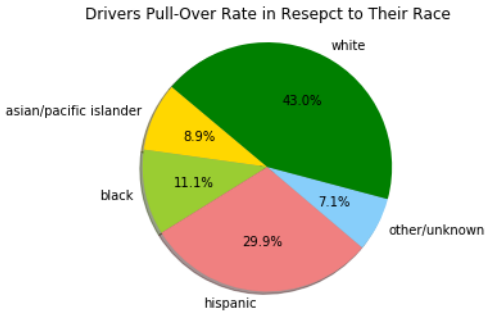
As we can see, our data now is clean and ready to be used; however, in order to build and fit our classifiers and predictors, the non-numerical features need to be transformed. That means the values of a feature such as 'subject_sex' must be transformed from {male, female} to numerical values {0,1} and so on for the other categorical-valued features. The following table shows part of the dataset after the transformation.

ıbject_age	subject_race	subject_sex	arrest_made	citation_issued	search_conducted	reason_for_stop
24.0	4	1	0	1	0	5
42.0	4	1	0	0	0	5
29.0	0	1	0	0	0	5
23.0	4	1	0	1	0	5
35.0	2	1	0	1	0	3

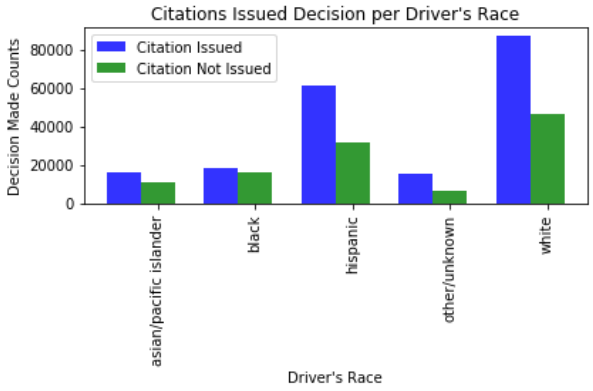
Data Exploring and Analysis

Statistical analysis with visualization

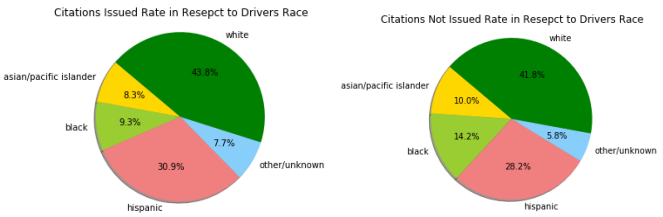




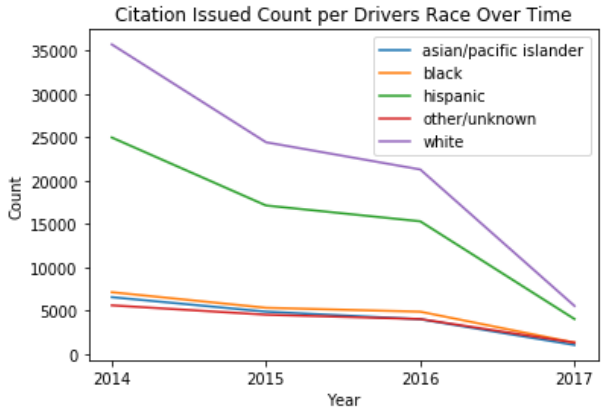
We begin our analysis by examining the demographic proportion of the people who have been pulled over. The two races of people that were most abundant were identified as white or Hispanic. People identified as white composed ~42% of the data set while Hispanic peoples represented ~30%. Taken together these two ethnicities account for ~72% of our total data set. While initially alarming because of the potentiality of a skewed data set, these values are consistent with the 2017 U.S. Census Bureau statistics of the ethnic makeup of San Diego. The census reports that 43.7% of the population was identified as white and 30% identified as Hispanic. This consistency between the proportion of pullovers and the actual demographics of the city validate the data set under examination. People identified as Black, Asian/Pacific islander or other/unknown comprised the rest of the data set, representing 11%, 8.8% and 7%.

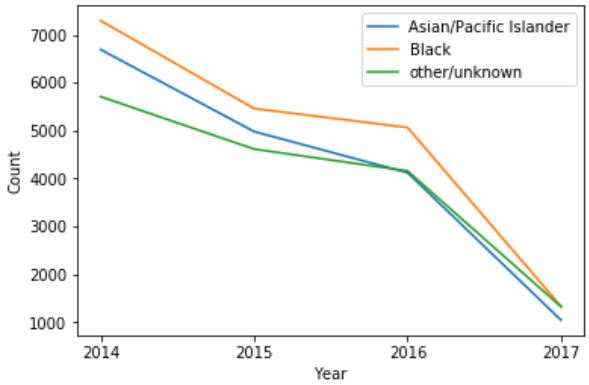


Next, we begin to assess the number of citations given versus not given in respect to race. People identified as white, hispanic or other/unknown were issued ~30%, 32%, 41% respectively more citations than not. The difference in issuing a citation versus not was minimal in the stops involving Black and asian/pacific peoples amounting to 8% and 18% respectively. All ethnicities received citations at a higher rate than not receiving them, but the people identified as white or hispanic had the largest difference between receiving citations out of all the demographics considered. This could be an artifact of the demographic composition of the data considering there is significantly more of these ethnicities as opposed to others.



The relative proportions of citations issued and not issued is conserved amongst all the ethnicities as the proportions are very similar between the two events. However, people classified as black did comprise 14.2% of people who did not receive a citation which is ~5% more than the fraction of they represent of people who did receive a citation. Aside from this instance, there does not seem to be a bias regarding receiving a citation





In the above figure the number of citations received in regard to race over a 3-year span is shown. The overall trend of citations received is decreasing for all ethnicities. People identified as white or hispanic seem to have the largest decrease in citations issued however this can begin to be explained by observing the share of the overall populations that they comprise. They make up most of the populations and so they will most likely receive a larger proportion of the citations. Consequently, when there is a decrease in the number of citations issued the drop rate will be more pronounced in these populations.

Machine Learning Models

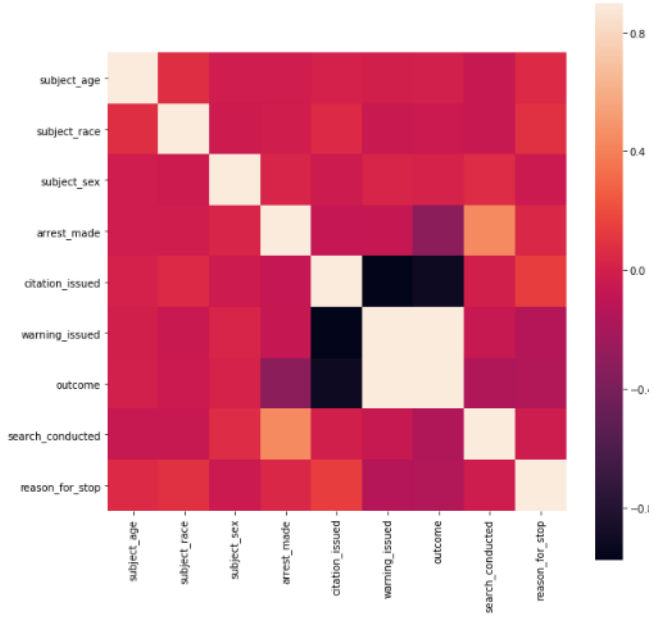
Building the Machine Learning Models

Before getting into machine learning and building our learning models, in addition to cleaning and transforming the data, still our dataset has large number of features. We need to reduce this number by using and applying different approaches available such as feature subset selection, sampling, dimensionality reduction (PCA), or features ranking approach. So, we will use

feature ranking and selection using forests of trees.

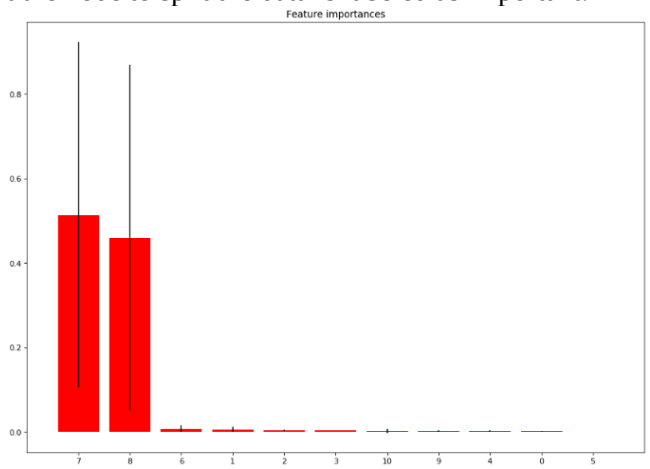
Features Ranking and Selection

First, let us see the correlation matrix of the features to understand how they are correlated to each other.



Here a correlation heat map is shown to illustrate the relationship between the features in the data set. For most of the features, there is no correlation but 'outcome' and 'warning_issued' have a strong positive correlation. While initially this correlation seems telling, upon closer examination the nature of this relationship does not provide insight to the data. The two possible outcomes are 'citation' or 'warning' so the outcome feature is just another way of coding the 'warning_issued' feature. Having both features is redundant to the data set and one of them can be removed. The other features that have a potentially interesting relationship is the 'citation_issued' and the 'warning_issued'. These two features have a strong negative correlation implying that they have a inverse relationship. Again, the strength of correlation between these features as misleading as the lack of a citation issue is the affirmation of a warning issue and vice versa. These two features are the converse of one another and resultantly introduce noise to our data set. One of these features can be dropped from our data. Although the correlation study did not provide and clues as to the relationships between the data, it did serve as a important feature reduction tool by providing two features to discard of, effectively reducing the problem space.

Second, let us run forests of trees algorithm and try all the possible outcomes from each feature regarding to the selector feature, 'citation_issued', and get a ranking of the features' importance. The quantitative measure used to assess the importance of a feature is the Gini impurity. The Gini impurity assess how well a split parsed the data into different classes by taking the relative count of each class into consideration before the split and comparing the resulting gini impurity after the split. If the split resulted in a child node that is largely or completely comprised of one class, then the feature used at the node to split the data is labeled as important.



FEATURE	INDEX	FEATURE IMPORTANCE
WARNING_ISSUED	10	57%
OUTCOME	5	40%
ARREST_MADE	4	0.9%
TIME	3	0.6%
SERVICE_AREA	1	0.3%
SUBJECT_AGE	2	0.3%
REASON_FOR_STOP	6	0.2%
SEARCH_CONDUCTED	7	0.2%
SUBJECT_RACE	8	0.1%
DATE	0	0.1%
SUBJECT_SEX	9	0.04%

Shown above is a bar graph expressing the importance of the features in classifying whether an instance will receive a citation or not and the boxplot shows the percentage of importance along with what index labels which feature. The features 'warning_issued' and 'outcome' were most useful in predicting whether a person will receive a citation or not. However, these two features are not helpful in our attempt to build a citation predictor as these features themselves are the labels for our predictions.

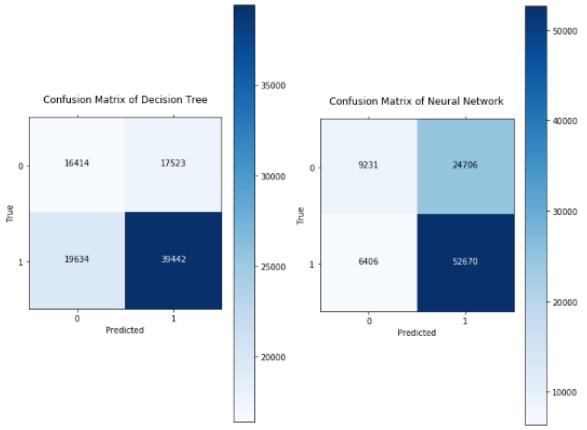
Machine Learning Methods

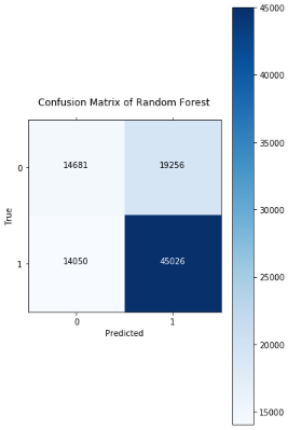
1. **Neural Network**
2. **Decision Tree**
3. **Random Forest**

It was concluded that the KNN algorithm would not be useful in attempting to predict if someone receive a citation or not. The data set was not comprised of values that consisted of measurements that describe the characteristic of the drivers. Before processing, the data was categorical and resultantly the values represented these categories. Distance between instances would not provide any insight. The Bernoulli Naive Bayes algorithm was not used since not all features were binary and the Gaussian Naive Bayes was not used because the data was not comprised of continuous values.

Performance Evaluation and Results

Model	Precision	Recall	f1-Score	Support
Artificial Neural Network	0.65	0.66	0.57	93013
Decision Tree	0.61	0.60	0.60	93013
Random Forest	0.63	0.64	0.64	93013





The artificial neural network, decision tree and the random forests all performed similarly with an average precision of 65%, 61% and 63% respectively. Similarly, the recall was similar amongst all models with 66%, 60% and 64% respectively. The artificial neural network had a bias in its predictions evidenced by 91% of the predictions it made were a citation being issued. However, the precision of predicting if a driver will be issued a citation is consistent with the overall precision at ~62%.

All models predicted most of the instances to receive citations.

	Class 0	Class 1
Ground Truth	36%	64%
ANN	7%	93%
Decision Tree	39%	61%
Random Forest	32%	68%

Shown above is the relative proportion of predictions from the three models. The Decision Tree and the Random Forest were able to maintain the correct proportions of the data when compared to the Artificial Neural Network. In the test labels, ~36% of the labels were not issued citations while ~64% were. This proportion was reflected in the Decision Tree and Random Forest models. ~62% of the Decision Tree's predictions were 'citation issued' while ~39% where 'citation not issued'. The Random Forest performed similarly as ~68% of its predictions were 'citation issued' and ~32% were 'citation not issued'. The predictions of the Artificial Neural Network were ~94%

‘citations issued’ and ~6% ‘citations not issued’.

CONCLUSION

While the model does not have a high precision or recall it is higher than 50% suggesting that the models are not assigning labels by chance. There must be an underlying structure to the data that can provide some implication to the whether a citation will be issued or not.

The confusion matrices show that all three models performed better at predicting ‘citation issued however this could be a result of the training data as the bulk of the data comprised of ‘citations issued. This is a direct result of the nature of the dataset as most of the dataset was comprised of drivers that were issued citations. An approach to amend the skewness could be the collection of more data and the addition of other features. Further studies should be conducted to discover the underlying variables.

ACKNOWLEDGEMENT

This paper describes a project that has been done at SDSU in the department of Computer Science for CS653 class (Data Mining and Knowledge). We are thankful to Dr. Xiaobai Liu for his guidance and cooperation.

References

[1] Kenneth Dowler. 2003.'Media consumption and public attitudes toward crime and justice the relationship between fear of crime punitive attitudes and perceived police effectiveness'. Journal of Criminal Justice and Popular Culture, 10(2) (2003) 109-126

[2] Shaw, T. M., & United States. (2015). The Ferguson report: Department of Justice investigation of the Ferguson Police Department.

[3] Lynn Langton and Matthew Durose. Police behavior during traffic and street stop, 2011. Technical report U.S. Department of Justice 2013.

[4] https://openpolicing.stanford.edu/

[5]A large-scale analysis of racial disparities in police stops across the United States. Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shr and Sharad Goel. March 13, 2019