Executive Summary

In this report, we set out to determine the factors that influence how people charged with possessing a small amount of marijuana are treated after their arrests. Specifically, we define their treatment as either being taken to court or detained/jailed. By performing statistical tests of significance and comparing logistic models, we decided that a marijuana arrestee's race, state of employment, age, the number of databases in which they appear, citizenship status, and the time period in which they were charged are significant predictors of the probability of being held by law enforcement. There were several interactions, including those between race and year, race and age, year and age, and between citizen and databases that are important in determining the probability of being held as well. Holding all other predictors constant, we have made the following conclusions: blacks are held more often than whites on average. When either age or number of databases increases, the probability of being held increases, more drastically for the latter. Unemployed arrestees are more likely to be held than employed arrestees, in much the same way that non-citizens are more likely to be held than citizens. The time of arrest also plays a role: people arrested between 2001-2003 were more likely to be held than those arrested between 2004-2006. Armed with this information, we make implications about our justice system and suggest avenues for future research.

Introduction

The legality of marijuana use has sparked intense debate in the United States, especially in the 21st century. Many Americans have been charged for possessing this drug, even in small quantities, but they have not all received equal treatment from law enforcement. Of course, this is sometimes due to the amount of marijuana that they have on hand, but controlling for this can help detect personal attributes that affect how drug users are treated. The data set provided controls for the violation's severity, and it includes nine predictor variables. The most pertinent among them are race, sex, the existence of traffic violations, regions of origin, employment, citizenship, and age. We also considered additional predictors such as the year the person was charged and the number of databases in which he/she appeared to determine if attitudes of law enforcement toward minorities have changed with time or if previously recorded police interactions impact how marijuana offenders are treated.

The goal of this study is to determine which combination of attributes leads to the highest probability of being detained or jailed in order to highlight any biases among U.S. law enforcement officers. In addition, these determinants can allow lawmakers to pass or revise legislation to mitigate some of the determinants (e.g., race, gender, age, etc.) that result in people being held in jail at disproportionate rates. This legislation could potentially prevent their lives from being upended and their reputations from being sullied over an illegitimate detainment or jail sentence. In addition, improved legislation can help mitigate the effects of detainment on the communities that are disproportionately affected.

We took a deliberate approach to achieve this goal. Before analyzing the data, we pre-processed it, which involved cleaning it. First, we removed any observations with missing predictor values and any observations with at least one predictor value outside the specified or reasonable range for that predictor. We then transformed some of the predictors, in some cases preserving their data type and in others converting them from continuous to categorical.

In R, we constructed two-way tables for the categorical predictors and obtained summary statistics for the numerical predictors. We ran chi-square tests for association on all the predictors, noting which were significant. Using only the four significant categorical predictors, we constructed spine plots reflecting the empirical probabilities that a person with a given attribute or characteristic was detained or jailed. We then constructed slicing and dicing plots for the two numerical variables.

The next important step was model selection. We began by ensuring that there was no collinearity between any two of the predictors. Next, we considered all possible logistic models with the six significant predictors and interactions between them, choosing the model with the lowest AIC. To confirm that the interaction terms that appeared in our model were consistent with the data, we produced interaction plots. Using our chosen model, we computed fitted probabilities of being held and plotted them against each of the two numerical variables, using different lines for each value of a given categorical variable.

To diagnose our model, we ran a goodness of fit test between our chosen model and the saturated model. We decided to classify an observation as "held" if the fitted probability was greater than 0.5 and "not held" if it was less than or equal to 0.5. With the aid of R, we created a classification table to catalog our correct and incorrect classifications and plotted a Receiving Operator Characteristic (ROC) curve.

Description of Subjects

The dataset initially contained 11 columns and 5,140 rows and it was arranged in wide format, each row in the data corresponding to a unique individual. The outcome variable in this dataset is "held," an indicator variable for which 1 corresponds to being held in jail while 0 corresponds to not being held in jail. Each observation has a unique identifier indicating that each row represents a unique person in the dataset. In addition, the predictors or explanatory variables present in the dataset are race, sex, prior traffic, region, employed, citizen, databases, year, and age.

After examining the variables in the dataset, the next step was cleaning the dataset. We removed the rows with no data so that we only considered candidates in the study that contained data for all the predictors. For example, if the person did not have data for race, they were omitted from the data analysis.

The second step for cleaning the data was ensuring that the values for each predictor were well-defined and that there were no extreme values. If there was a value for a predictor that wasn't well defined, the entire observation (row) was removed from the dataset. For example, if someone was cataloged as having been arrested in 2000, then this entire observation (row) was

removed from the data. This is an example of a potential mistake in the dataset since the year of arrest should only cover the time period from 2001 to 2006. After going through all these steps for the potential predictors, we also removed duplicate observations. Even if the person had identical values for many of their predictors, we were able to determine which observations were unique since the data contained the unique ID number for each participant.

The next step for data cleaning was re-coding the variables in a way that would be meaningful to our data analysis. For the predictor databases, the number of people that appeared in six databases was 9, a relatively small number. To remedy this, we collapsed the category so that the largest level, 5, represents that the person appeared in at least 5 databases. We also re-encoded the year to reduce the number of factors and simplify our interpretation. The categories were changed to represent three-year increments where arrests before 2004 are in the level "Before 2004," and arrests during and after 2004 are labeled "2004 or later." These are two main variables that benefit from being re-encoded.

We continued our data analysis by examining the characteristics of the subjects more closely. To begin, we constructed two-way tables to determine the relative amount of observations in each category. We calculated and compiled the relative percentages of each level of the categorical variable for the outcome categories "held" and "not held" into a table. Table 1 contains a column for the p-value corresponding to a test of association for each categorical predictor. In order to look more carefully at a subset of the variables, we ran a Chi-square test on each variable to see if it was associated with the outcome variable (Held). The null hypothesis for the test was that the given predictor was not associated with being held in jail on marijuana charges. The alternative hypothesis was that there was an association between the predictor and being held in jail on marijuana charges. In the last column of the table, we record the resulting p-value for the chi-square test. The critical value for most of these tests was 3.84, as this is the 95th percentile of a chi-square distribution with 2 - 1 = 1 degree of freedom. The two exceptions were the tests on the variables prior traffic and region with critical values of 5.99 and 7.81, respectively. The significant test statistics (with a p-value less than $\alpha = 0.05$) are highlighted in light gray in Table 1. The categorical variables that were significantly associated with being held in jail due to marijuana charges were race, employment status, citizenship status, and year. Since these categorical variables are significant, we will consider them for the logistic regression model.

Table 1: Descriptive Statistics for Predictors of Being Held in Jail					
Predictor	Categories	Total	Held in Jail, N (Row %)	Not Held in Jail, N (Row %)	Chi-Square Test P-Value
Race	White	3864	549, (14.21)	3315, (85.79)	< 2.2*10^(-16)
	Black	1269	325, (25.61)	944, (74.39)	
Sex	Male	4694	813, (17.32)	3881, (82.68)	0.07854
	Female	439	61, (13.9)	378, (86.1)	
Prior Traffic	0	2038	347, (17.03)	1691, (82.97)	0.9892
	1	1747	299, (17.12)	1448, (82.88)	
	2	1348	228, (16.91)	1120, (83.09)	
Region	North	1517	279, (18.39)	1238, (81.61)	0.3878
	South	1036	172, (16.6)	864, (83.4)	
	East	1054	169, (16.03)	885, (83.97)	
	West	1526	254, (16.64)	1272, (83.36)	
Employed	Yes	4034	531, (13.16)	3503, (86.84)	< 2.2*10^(-16)
	No	1099	343, (31.21)	756, (68.79)	
Citizen	Yes	4384	671, (15.31)	3713, (84.69)	3.127*10^(-15)
	No	749	203, (27.1)	546, (72.9)	
Year (Time Period)	Before 2004	2429	445, (18.32)	1984, (81.68)	0.0215
	2004 or Later	2704	429, (15.87)	2275, (84.13)	

Race

This variable indicates whether a person identifies as black or white. The dataset appears to have more than three times as many white people as black people. Also, the relative percentage of black people held in jail is much larger than the corresponding percentage of white people. The test of association indicates that being held in jail is associated with race, so we decided to consider this predictor when choosing between logistic regression models.

Sex

The variable sex, which encodes whether each person in the dataset is male or female, did not clear the significance threshold. From the relative percentages, it does not appear that males and females in the dataset are detained for marijuana charges at higher rates. However, this dataset has a disproportionate amount of men, as there are more than 10 times as many men as women. Yet, since the p-value is just above 0.05, we cannot reject the claim that there is no association between sex and being detained or jailed.

Prior Traffic

Since it appears that around 40% of the people in the dataset did not have any prior traffic convictions, this group is overrepresented in the data set. Still, there is no association between the number of prior traffic convictions and the relative percentage of people held in jail.

Regardless of the number of prior traffic convictions, it appears that around 17% of arrestees in each category were held in jail.

Region

This dataset contains observations from the four possible regions; “North,” “South,” “East,” and “West.” This predictor is not associated with being held in jail since the p-value of 0.3878 for the test of association is higher than the threshold alpha level of 0.05. The Northern region contains the highest percentage of offenders held in prison at 18%, whereas the other three regions have an approximately equal percentage of 16%. This predictor is not significant in determining if a person will be held in prison.

Employed

For the variable employed, if a person was working either a part-time or full-time job, they were considered employed (“Yes”). On the other hand, if the person was a student, retiree, or unemployed, they were considered not employed (“No”). Given that 31.21% of all unemployed people in the study were held in jail compared to 13.16% of employed people, there appears to be an association. This association is confirmed by the p-value, which is lower than the critical alpha level, leading us to reject the hypothesis that there is no association between the two variables.

Citizen

The citizen variable codes “Yes” for people who are considered US citizens, and all others are encoded as “No” for this predictor. 27.1% of all non-US citizens were held in jail, whereas only 15.31% of US citizens were held in jail, indicating that there may be an association between detainment status and citizen status. The test of association allows us to confirm that there is indeed an association between being held in jail and an arrestee’s US citizenship status.

Year

The last categorical variable is the year in which each person was arrested, and it has been transformed into two time periods. Marijuana legislation in Connecticut, New Mexico, Vermont, and Wyoming was introduced in 2003. So, it is possible that existing legal or police guidelines on arresting people for marijuana possession charges either changed over time or that other new legislation was passed. In the dataset, those arrested between 2001 and 2003 are combined into one group, and those arrested between 2004 and 2006 are placed into another group. This transformed variable, year, is associated with being held in jail; and the p-value is 2.1%, less than the 5% critical level.

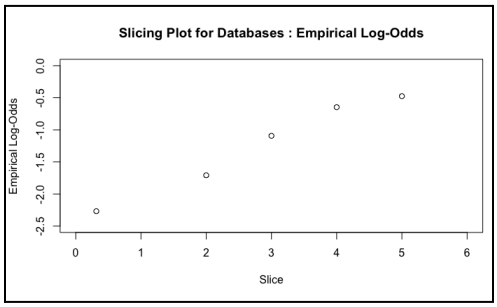
Table 2: Descriptive Statistics

Predictor	Databases	Age
Minimum	0	13
1st Quartile	0	19
Median	1	22
Mean	1.644	24.86
3rd Quartile	3	28
Maximum	6	67
LR Test (P-value)	< 2.2*10^(-16)	0.001631

The numerical variables must be analyzed according to their distributions, so the five-number summary statistics and test results for the numerical variables can be found above in Table 2. The p-value in the bottom row of Table 2 is a likelihood ratio test of the model including the predictor and the intercept-only model. Since both of the p-values are less than the threshold of 0.05, we can conclude that they are good predictors of being held in jail.

Databases

The slicing plot for the number of databases in which an arrestee appears displays a moderate linear relationship between the empirical log-odds for each average slice value for the database variable. Thus, it is reasonable to say that there is a positive linear relationship between the log-odds of being held and the number of databases in which a person appears. We choose to treat the number of police databases that a person appears in as a numerical variable and find that it is a significant predictor of the log odds of being held in jail.



Age

The last variable considered for the model is the age (in years) at which a person is charged with marijuana possession, and we treat it as numerical. The slicing plot for the average age in an age group also indicates that there is roughly a linear relationship. Therefore, as age increases, the log-odds of being held increase as well. This relationship allows us to infer that there is a linear effect on the log-odds of being held based on the age of the person. Age is significant since the p-value for the likelihood ratio comparing a model with age and an intercept-only model is 0.0016. This value is much smaller than the critical value, so it is a significant predictor to be considered for the model.

Having obtained this subset of significant predictors, we proceed to construct a correlation coefficient matrix to ensure there is no collinearity amongst the potential predictors. The correlation coefficients can be found below in Table 3. None of the correlations is above 0.25 for any 2 unique predictors. Thus, we are able to consider all of these predictors.

Table 3: Correlation Matrix

Predictor	Race	Employed	Citizen	Databases	Year	Age
Race	1.000	0.107	0.219	-0.171	-0.004	-0.068
Employed	0.107	1.000	0.076	-0.245	-0.033	-0.118
Citizen	0.219	0.076	1.000	-0.033	-0.159	-0.074
Databases	-0.171	-0.245	-0.033	1.000	0.027	0.136
Year	-0.004	-0.033	-0.159	0.027	1.000	0.017
Age	-0.068	-0.118	-0.074	0.136	0.017	1.000

Results

Before selecting the predictors to be fitted in a logistic regression model, we remove the variables deemed insignificant by the chi-square test we conducted earlier and focus on the predictors that have a significant association with the response variable "held." This includes four categorical variables, Race, Employed, Citizen, and Year, and two continuous variables, Databases and Age.

As discussed previously, we have transformed the variable Year into a categorical variable with two factor levels – "before 2004" and "2004 or later,". Since the empirical probabilities of being held across years, holding other variables constant, do not form a linear relationship, the model is simpler and more interpretable. Additionally, since there are very few observations (~0.18%) with Databases equal to 6, it would be beneficial to transform all entries with Databases = 6 into Databases = 5 to ensure that the few edge cases would not have a large impact on our model coefficients. We then plotted the empirical log odds on a slicing-dicing plot across different values of databases (as shown on the previous page), revealing a linear trend. This indicates that we can treat Databases as a continuous variable when fitting logistic models.

Race

	Held	Pr(Held)	Not Held	Pr(Not Held)	Total
Black	325	0.2561072	944	0.7438928	1269
White	549	0.1420807	3315	0.8579193	3864
Total	874	1.0000000	4259	1.0000000	5133

Year

	Held	Pr(Held)	Not Held	Pr(Not Held)	Total
2004 or later	429	0.1586538	2275	0.8413462	2704
Before 2004	445	0.1832030	1984	0.8167970	2429
Total	874	1.0000000	4259	1.0000000	5133

Employed

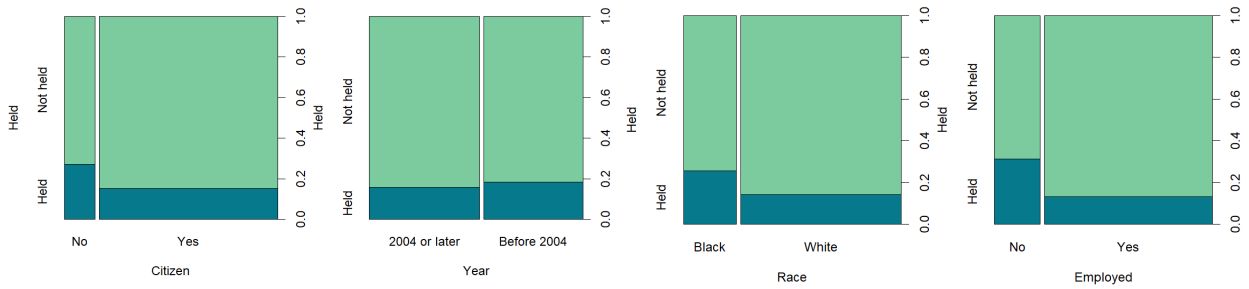
	Held	Pr(Held)	Not Held	Pr(Not Held)	Total
No	343	0.3121019	756	0.6878981	1099
Yes	531	0.1316311	3503	0.8683689	4034
Total	874	1.0000000	4259	1.0000000	5133

Databases & Age

```
summary(policing2$databases)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.642   3.000   5.000

summary(policing2$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   19.00   22.00   24.86   28.00   67.00
```

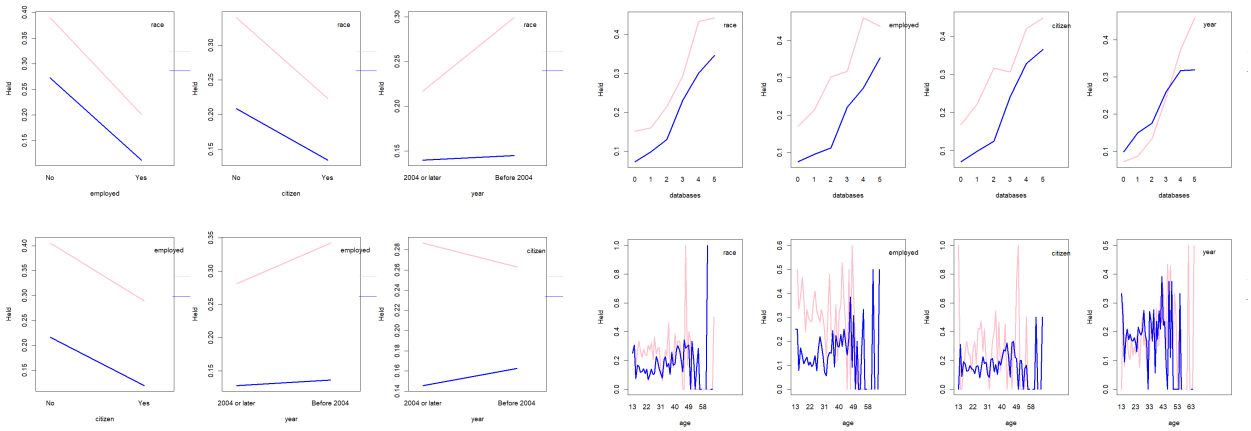
Citizen

	Held	Pr(Held)	Not Held	Pr(Not Held)	Total
No	203	0.2710280	546	0.7289720	749
Yes	671	0.1530566	3713	0.8469434	4384
Total	874	1.0000000	4259	1.0000000	5133



We have constructed a two-way table and a corresponding mosaic plot for each categorical variable. The mosaic plots with citizenship status, race, and employment status on the horizontal all have similar shapes in that the minority group for each has a higher empirical probability of being held than the majority group. That is, noncitizens, blacks, and unemployed people had a higher probability of being held than citizens, whites, and employed people, respectively. The mosaic plot with year on the horizontal shows a more even split of observations, with about 47% arrested before 2004 and 53% arrested in 2004 or later. It also shows similar probabilities of being held for both time periods. Those arrested before 2004 were about 2.5 percentage points more likely to be held than those in 2004 or later.

Since the response variable Held is categorical with binary outcomes, fitting a logistic regression model to predict the probabilities is a sensible choice. Since there are a maximum of six main effects included in the model, we decided to implement an algorithm in R to perform model selection using the best subset selection method. This algorithm considers all possible models that include combinations of main effects and two-way interaction terms (subjected to the hierarchy principle). Then the algorithm stores each model and its corresponding AIC and BIC values in a data frame, and it returns the five models that give the lowest AIC and BIC values, respectively. The output of the algorithm is summarized in the appendix.



Next, we identified the significant interaction terms in the model by creating interaction plots for each pair of predictors. From the interactions plots above, we concluded that interactions between citizen and employment status, race and employment status, race and

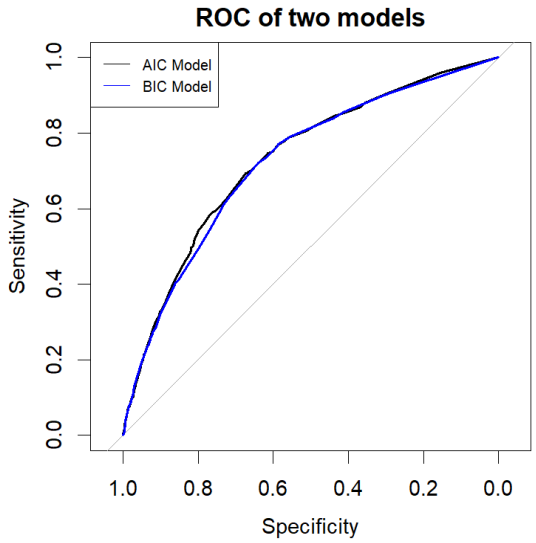
citizen, and between database and employment status are not significant since the pair of curves are parallel or close to parallel in these respective plots. The best subset model with the lowest AIC value shown below does not contain any of these interaction pairs, which validates our choice of interaction terms.

We refitted the logistic model using the aforementioned selected variables and interaction terms, yielding the following general model:

$$ln\left(\frac{Pr(held)}{1-Pr(held)}\right) = \beta_{race}race + \beta_{employed}employed + \beta_{citizen}citizen + \beta_{databases}databases + \beta_{year}year + \beta_{age}age + \beta_{race:year}race(year) + \beta_{race:age}race(age) + \beta_{employed:year}employed(year) + \beta_{employed:age}employed(age) + \beta_{citizen:databases}citizen(databases) + \beta_{citizen:year}citizen(year) + \beta_{databases:year}databases(year) + \beta_{year:age}year(age)$$

Variable Encoding:

Race	White = 1 Black = 0
Employed	Yes = 1 No = 0
Citizen	Yes = 1 No = 0
Year	Before 2004 = 1 2004 or Later = 0




```
Call:
glm(formula = held ~ race + employed + citizen + databases +
    year + age + race:year + race:age + employed:year + employed:age +
    citizen:databases + citizen:year + databases:year + year:age,
    family = binomial, data = policing2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5731  -0.6325  -0.4281  -0.3124   2.5271

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.323906   0.386659  -0.838 0.402197
racewhite                   -0.900623   0.289514  -3.111 0.001866 **
employedYes                 -1.126999   0.287752  -3.917 8.98e-05 ***
citizenYes                  -1.132989   0.213905  -5.297 1.18e-07 ***
databases                    0.353652   0.065681   5.384 7.27e-08 ***
yearBefore 2004              1.119707   0.360311   3.108 0.001886 **
age                         -0.031666   0.011628  -2.723 0.006467 **
racewhite:yearBefore 2004   -0.568101   0.173792  -3.269 0.001080 **
racewhite:age                0.032145   0.010208   3.149 0.001639 **
employedYes:yearBefore 2004 -0.368213   0.171079  -2.152 0.031374 *
employedYes:age              0.021749   0.009822   2.214 0.026808 *
citizenYes:databases         0.137768   0.062702   2.197 0.028009 *
citizenYes:yearBefore 2004   0.417789   0.209670   1.993 0.046305 *
databases:yearBefore 2004   -0.182453   0.053708  -3.397 0.000681 ***
yearBefore 2004:age         -0.014007   0.009357  -1.497 0.134391
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4684.5  on 5132  degrees of freedom
Residual deviance: 4165.6  on 5118  degrees of freedom
AIC: 4195.6

Number of Fisher Scoring iterations: 5
```

Confusion matrices:

Held.Hat.AIC	0	1
0	4212	47
1	815	9

Held.hat.Bic	0	1
0	4238	21
1	842	32

We employed a three-step approach to assess the fit of our model. First, we conducted a chi-square goodness of fit test on both the lowest AIC and BIC models. Both the lowest AIC and BIC models yield p-values very close to 1, with residual deviances of 4165.6 and 4211.4 on 5118 and 5127 degrees of freedom, respectively. Both models thus fit the observations closely.

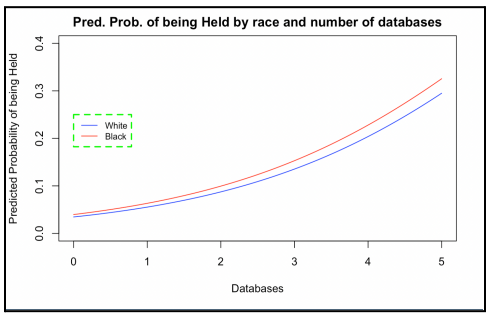
Next, we produced a confusion matrix for each model, which uses a cutoff value of 0.5 for the predicted probability of being held, as shown above. Further, we calculated the error rate

of the lowest AIC model to be 16.79%, meaning the model makes accurate predictions on the training data around 83.21% of the time; for the lowest BIC model, the error rate is 16.81%, so the model makes accurate predictions on the training data around 83.19% of the time. In particular, the two models produce small false positive (type I error) rates of 1.10% and 0.49%, respectively, but large false negative (type II error) rates of 93.25% and 96.34% respectively. The large false negative rates show there is room for improvement; one possible solution is to lower our cutoff probability value from 0.5 in order to achieve a better tradeoff between the two types of errors.

Lastly, we constructed ROC curves for both the lowest AIC and lowest BIC models. The AUC value of the lowest AIC model is 0.7336, while that of the lowest BIC model is 0.7241. We conclude that the lowest AIC model, which we selected as our final model, displays better predictive performance.

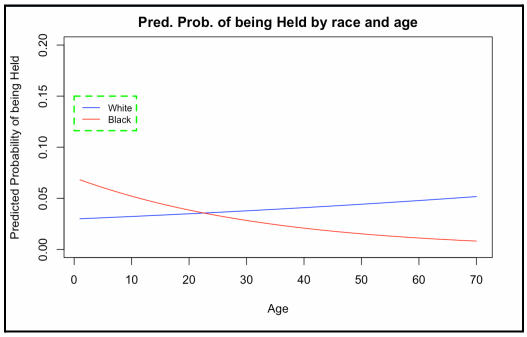
Using the R summary output for our final model, we created a summary table with parameter estimates, odds ratios, odds ratio confidence intervals, and p-values for each variable in the model. When calculating the odds ratio, we are using the default value of 1 for all the categorical variables (Race, Employed, Citizen, Year), the value 19 for Age, and a value of 1 for Databases, when the variable is not the subject of the odds ratio. Some of the significant results from the table will be discussed in the next section.

Main Effects	Estimate	Odds Ratio	Odds Ratio CI	p-value
Race	-0.9006	0.4240	(0.3449, 0.5031)	0.0019
Employed	-1.1270	0.3389	(0.2589, 0.4190)	<0.0001
Citizen	-1.1330	0.5613	(0.4690, 0.6536)	<0.0001
Databases	0.3537	1.3620	(0.9532, 1.9462)	<0.0001
Year	1.1197	1.1649	(1.0905, 1.2392)	0.0019
Age	-0.0317	1.0083	(0.9304, 1.0927)	0.0065



In producing the probability plots with the number of databases on the horizontal axis, we set all categorical variables not being examined to their most common values (employment = employed, citizenship = citizen, time period = after 2004, and age = 19). The probability plots with age on the horizontal axis are constructed similarly, with databases equal to its most common value, 0. As the number of databases in which an offender appears increases, the probability the offender has been

arrested increases for both white (0.035 for 0 databases vs. 0.295 for 5) and black (0.040 for 0 databases vs. 0.326 for 5) races, with the probabilities for black races being slightly higher for all database values. As the age of an offender increases, the probability that he has been detained increases if he is white, whereas the probability decreases if he is black. The probabilities of being held are the same for both races (0.04) between ages 22 and 23.



Discussion

The significant predictors in the model include race, employment status, citizenship status, databases, year, and age, and their combined main effects along with the significant interaction effects result in a well-fitted model. Therefore, it is important to understand the effect and direction of each of the predictors in the model. In order to contextualize and give meaning to the effect and direction of each predictor, we will use odds ratios.

We can analyze the strength and direction of the effect of the predictor employed using the odds ratio comparing an unemployed and employed 19-year-old who was charged with possession of marijuana between 2001 to 2004. The odds of an employed person being held in jail are 66.1% lower than the odds of being held in jail for someone who is unemployed. Both the age of a person and the time period in which he was caught with marijuana affects his chance of being held. An example of this can be illustrated by the fact that the odds of an employed 30-year-old person being held in jail is 57% less than an unemployed 30-year-old person being held in jail during the time period from 2001 to 2004. So, as a person gets older the chance of them being held in jail is less impacted by their employment status.

The predictor citizen has both main effects in the model along with interaction effects. From the model, it appears that the effect of being a citizen on the odds of being held in prison is different depending on when a person was charged along with the number of databases the person can be found in. For example, the odds of a 19-year-old citizen of the United States found in a database being held is 43.9% less than the odds of being held for someone of the exact profile, but they are not a citizen during the period from 2001 to 2004 who also appeared in one database. The interaction terms are important for this predictor as well since it can be seen in the model that the number of databases a person appears in along with the time period have an effect on a citizen or non-US citizen being held. Using the same example from before with one modification that a person appears in two databases results in an odds ratio of 35.6%. This is less than the odds ratio from before, and this indicates that being found in a police database decreases the impact of citizen status on the odds of being held in jail for marijuana.

The predictor databases has both a main effect along with the interaction effects with the citizen and time period predictors. For every additional database that a 19-year US citizen can be

found in between the time period of 2001 to 2004, the odds of being held in jail are 36.2% higher. So, if a person can be found in more databases they have a higher chance of being held in jail due to possession of marijuana.

Another predictor that is significant in determining if a person will be held in jail as a result of marijuana possession is the time period in which he was charged. This indicates that perceptions of how police officers should address people with marijuana charges have changed over time. For example, the odds of being held in jail for a 19-year-old employed US citizen is 82% higher compared to a person with the same profile from the period from 2004-2006. Thus, the odds of being detained for marijuana possession appear to have generally decreased from 2003 to 2004.

In addition, the predictor age impacts whether a person is detained, and other predictors such as race, employment status, and time period change the impact of age on the manner in which an arrestee is treated. For each additional 1-unit increase in age, the odds of being held are 5.4% lower for a white employed person arrested between 2001 and 2004. As a person with the exact same profile in the previous increases by 10 years, the odds of being held is 42.9% lower. Therefore, it appears that younger people have a higher chance of being detained if they are caught with marijuana.

For the variable race, the odds of being held in jail for a 19-year-old person between 2001 to 2004 is 57.5% less than that of a 19-year-old black person charged with marijuana possession during the same time period. Conversely, this means that the odds of a 19-year-old black person being held in jail are 135.8% higher than the odds of being held in jail for a 19-year-old white person. This percentage indicates that there is a discrepancy between the odds of a black person being held compared to that of a white person in the same age group and during the same time period. So, it is clear that black people are held at higher rates compared to white people for marijuana possession. The odds ratio is much higher here, and this is evidence that there is indeed racial profiling occurring. Even after taking into account the other variables that impact if a person is detained based on their race, there is still evidence that black people are detained for marijuana at higher rates than white people.

Thus, this model provides evidence that racial profiling is occurring when people are detained at higher rates due to their race. This phenomenon can be further analyzed in order to determine its impact on communities of color. This research can be extended by examining a larger set of predictors, a longer time period, and analyzing models with other tools. Other predictors that may impact the rate at which people are detained for marijuana usage include income, the race of the police officer arresting the individual, the amount of police presence in the region, the time of day, and the political affiliation of the region. Given the recent legislation surrounding decriminalizing marijuana usage, it might be interesting to determine how rates of being detained have changed over time with respect to variables such as race.

Appendix

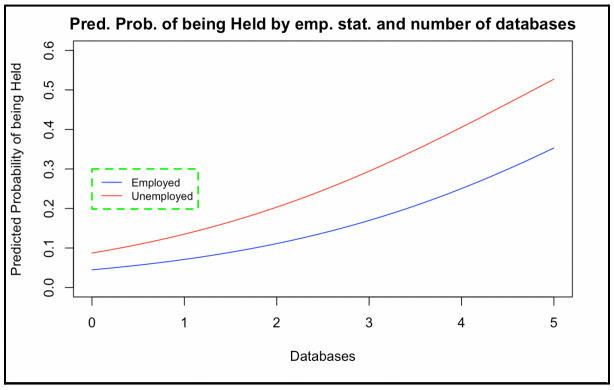
I. Model Selection

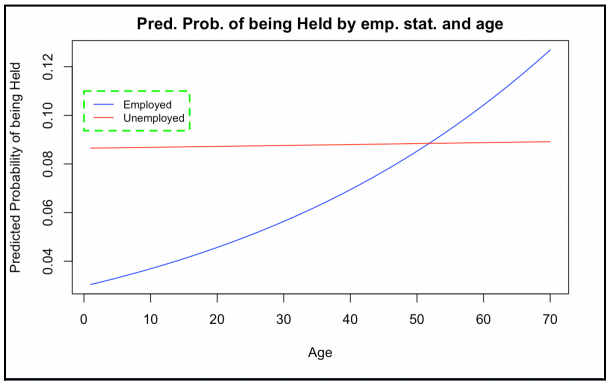
Top 5 AIC Models			
Iteration	Model	AIC	BIC
29755	held ~ race + employed + citizen + databases + year + age + race:year + race:age + employed:year + employed:age + citizen:databases + citizen:year + databases:year + year:age	4195.597	4293.748
23034	held ~ race + employed + citizen + databases + year + age + race:year + race:age + employed:year + employed:age + citizen:databases + citizen:year + databases:year	4195.842	4287.450
34962	held ~ race + employed + citizen + databases + year + age + race:year + race:age + employed:year + employed:age + citizen:databases + citizen:year + citizen:age + databases:year + year:age	4196.844	4301.540
34937	held ~ race + employed + citizen + databases + year + age + race:year + race:age + employed:databases + employed:year + employed:age + citizen:databases + citizen:year + databases:year + year:age	4197.147	4301.842
34964	held ~ race + employed + citizen + databases + year + age + race:year + race:age + employed:year + employed:age + citizen:databases + citizen:year + databases:year + databases:age + year:age	4197.180	4301.876

Top 5 BIC Models			
Iteration	Model	AIC	BIC
168	held ~ race + employed + citizen + databases + citizen:databases	4223.429	4262.690
2134	held ~ race + employed + citizen + databases + age + race:age	4220.704	4266.508
1147	held ~ race + employed + citizen + databases + year + race:year + databases:year	4215.042	4267.389
2168	held ~ race + employed + citizen + databases + age + race:age + citizen:databases	4215.258	4267.606
1117	held ~ race + employed + citizen + databases + year + databases:year	4222.348	4268.152

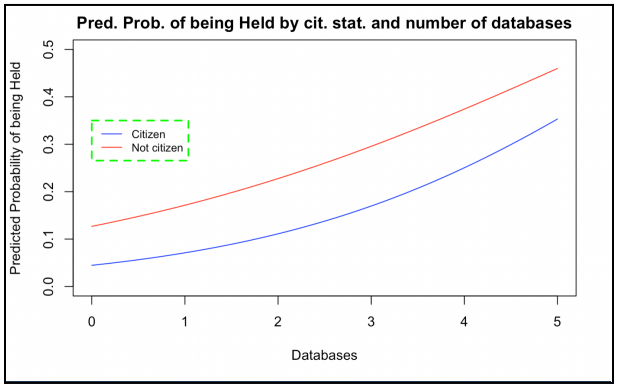
II. Probability plots with categorical variables other than race

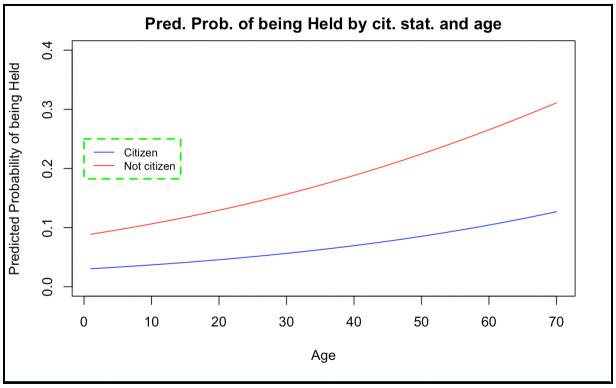
A. Employment status



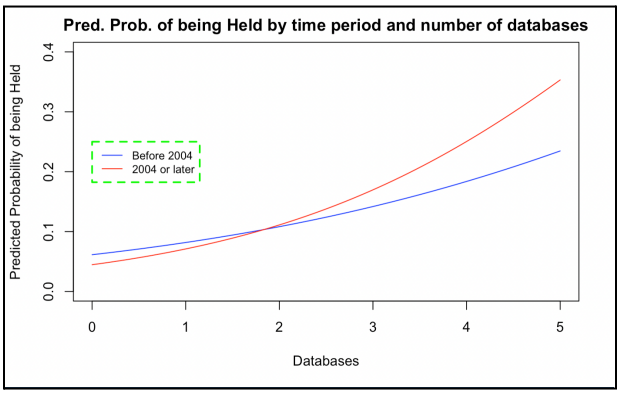


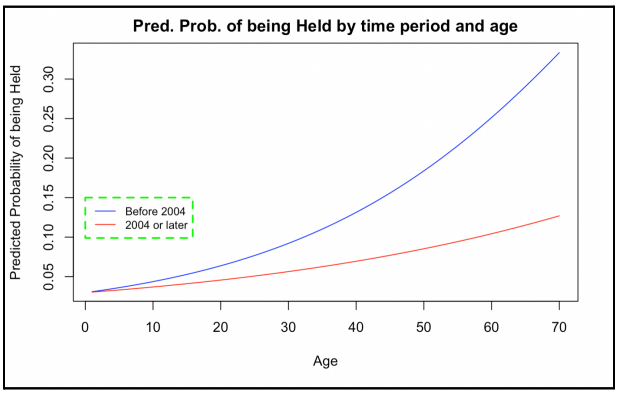
B. Citizenship status





C. Time period (year)





III. Predicted Probabilities

Databases	**0**	**1**	**2**	**3**	**4**	**5**
White	0.0346	0.0554	0.0874	0.1354	0.2038	0.2950
Black	0.0397	0.0633	0.0995	0.1530	0.2280	0.3255

(holding age = 19, employed = 1, citizen = 1, time period = after 2004)

Age	**15**	**20**	**35**	**50**	**65**	**70**
White	0.0335	0.0349	0.0393	0.0442	0.0497	0.0517
Black	0.0448	0.0385	0.0243	0.0153	0.0095	0.0082

(holding databases = 0, employed = 1, citizen = 1, time period = after 2004)