Analysis of Arrests for Marijuana Possession

1. Executive Summary

In terms of the issue of racism in the American criminal justice system, one can examine the multitude of factors that influence the decision of the police to grant release to arrestees and whether the ethnicity of the arrestees is a significant predictor. In this project, we hope to explore the following research questions: what are the factors that are relevant to the police's treatment of a person arrested for marijuana possession? What are the factors that can best predict the harsher treatment of being held in jail or the police station received by the arrestees? Is the race of the arrestee a significant predictor?

To answer these questions, we analyzed data collected from 5140 individuals who were arrested for possession of marijuana during the time period from 2001 to 2006 in the United States. A logistic model was fitted to predict the outcome of being held in detention or released.

From this model, we concluded that, when controlling for other variables, being African American, having appeared in police databases, being unemployed, and having no U.S. citizenship are associated with a higher probability of the arrestee being held in jail or the police station. In addition, the association between the arrestee's ethnicity and the probability of being held in jail, as well as the association between the arrestee's employment status and the probability of being held in jail, both depend on the age of the arrestee.

2. Introduction

According to Article 9 of the Universal Declaration of Human Rights, arbitrary arrest and detention are violations of fundamental human rights.[1] To define arbitrariness, the Human Rights Committee interprets that *remand in custody*–in layman's terms, being held in detention or prison waiting for their trials or sentences–must be necessary and reasonable.[2] In the United States, placing a person on remand is only reasonable if they pose a flight risk, might interfere with evidence, or repeat previous crimes.[3] Thus, it constitutes arbitrary detention when a person, who does not pose any of the said risks, is deprived of their liberty on grounds of ethnicity or race alone. In other words, racial profiling in terms of holding one in detention without a justifiable cause is a fundamental right violation.

To investigate the presence of such human rights violations in the American criminal justice system, this project analyzes data collected from 5140 individuals who were arrested for possession of a small amount of marijuana during the time period from 2001 to 2006 in a major American city. In order to determine the factors predictive of the treatment of being held in

---

[1] United Nations. Universal Declaration of Human Rights. 1948.

[2] High Commissioner for Human Rights, Centre for Human Rights, Geneva Staff. *Human rights in the administration of justice: A manual on human rights for judges, prosecutors and lawyers*. United Nations Publications, 2003.

[3] Hopkins, Brook, Chiraag Bains, and Colin Doyle. "Principles of pretrial release." *The Journal of Criminal Law and Criminology (1973-)* 108.4 (2018): 679-700.

prison or police station after arrest and investigate whether racial profiling is present, we use logistic regression analysis. We first conducted data cleaning by removing abnormal values, then we used two-way contingency tables to eliminate variables that do not have an association between the outcome variables. Then, we chose the base model using the trend test, association test, and slicing dicing plot to remove insignificant predictors. We then utilized the forward selection method with likelihood ratio tests to select models. Through examining Akaike information criterion (AIC) values, we determined the final model that could best predict the harsher treatment of being held at the jail or police station after arrest.

3. Description of Subjects

3.1 Data Cleaning

After making sure that there was no missing data, we checked each column for abnormal or out-of-range data. We found that for the *race* variable, there was an abnormal value "Gr"; for the *databases* variable, there was an abnormal data "33"; for *region*, there was an abnormal data "Purple"; for *citizen*, there was "Neverever"; for *year*, there was an abnormal "1215"; for *age*, there were "117" and "3". Upon deleting these values, we have a clean dataset with 5133 rows of data for analysis.

In addition, we combined databases 5 and 6 because database 6 only had nine values, which was significantly lower than the average number of values in the rest of the databases.

3.2 Descriptive Statistics

Among the ten variables, all are categorical except for *age*, which is a numerical variable. As **Table 1** demonstrates, regarding race, 75.3% of the dataset is white and 24.7% is black. With regard to sex, only 8.5% are female and 91.5% are male. 39.7% had no prior traffic convictions, 34.0% had one prior traffic conviction, and 26.3% had two or more prior traffic convictions. For the regional district of people who were arrested for marijuana possession, 20.5% were arrested in the east, 29.6% of the people were arrested in the north, 20.2% were arrested in the south, and 29.7% were arrested in the west. As for employment status, 78.6% were employed and 21.4% were unemployed. 85.4% of the arrestees were U.S. citizens and 14.6% were non-U.S. citizens. In terms of the number of police databases the arrestees appeared in, 35.3% had never appeared in any police record, 16.2% appeared in one database, 15.1% appeared in two databases, 18.4% appeared in three databases, 12.4% appeared in four databases, 2.4% appeared in five databases, and 0.2% appeared in six databases. Concerning the year of the arrest, 9.35% of people were arrested in 2001, 16.9% were arrested in 2002, 21.2% were arrested in 2003, 24.2% were arrested in 2004, 23.1% were arrested in 2005, and 5.3% were arrested in 2006.

Table 1: Descriptive statistics for held categorical predictors. Binary/Nominal variables highlighted in blue. Ordinal variables highlighted in orange.

Predictor	Categories	Total, N (%)	Held in a jail for bail hearing, N (%)	Summons to appear in court, N (%)	Chisq Test
Race of person who was arrested for possession of small amount of marijuana	White	3864 (75.3)	549 (14.2)	3315 (85.8)	$\chi2 = 87.915$, df=1, p-value < 2.2e-16
	Black	1269 (24.7)	325 (25.6)	944 (74.4)	
Sex of person who was arrested for possession of small amount of marijuana	Female	439 (8.5)	61 (13.9)	378 (86.1)	$\chi2 = 3.3329$, df=1, p-value = 0.06791
	Male	4694 (91.5)	813 (17.3)	3881 (82.7)	
Prior traffic convictions for person who was arrested for possession of small amount of marijuana	No prior traffic convictions	2038 (39.7)	347 (17.0)	1691 (83.0)	$\chi2 = 0.00492$, df=1, p-value = 0.9440792
	1 prior traffic conviction	1747 (34.0)	299 (17.2)	1448 (82.9)	
	2 or more prior traffic convictions	1348 (26.3)	228 (16.9)	1120 (83.1)	
Regional district of person who was arrested for possession of small amount of marijuana	East	1054 (20.5)	885 (84.0)	169 (16.0)	$\chi2 = 3.0248$, df=3, p-value =0.3878
	North	1517 (29.6)	1238 (81.6)	279 (18.4)	
	South	1036 (20.2)	864 (83.4)	172 (16.6)	
	West	1526 (29.7)	1272 (83.4)	254 (16.6)	
Employment status of person who was arrested for possession of small amount of marijuana	Employed (Yes)	4034 (78.6)	531 (13.2)	3503 (86.8)	$\chi2 = 199.11$, df=1, p-value < 2.2e-16
	Unemployed (No)	1099 (21.4)	343 (31.2)	756 (68.8)	
Citizenship of person who was arrested for possession of small amount of marijuana	US Citizen (Yes)	4384 (85.4)	671 (15.3)	3713 (84.7)	$\chi2 = 63.017$, df=1, p-value = 2.049e-15
	Non-US Citizen (No)	749 (14.6)	203 (27.1)	546 (72.9)	
Databases count indicating the number of police databases the person appeared in	0 Databases	1810 (35.3)	153 (8.5)	1657 (91.5)	$\chi2 = 319.7994$, df=1, p-value < 2.2e-16
	1 Databases	833 (16.2)	95 (11.4)	738 (88.6)	
	2 Databases	776 (15.1)	119 (15.3)	657 (84.7)	
	3 Databases	944 (18.4)	237 (25.1)	707 (74.9)	
	4 Databases	637 (12.4)	219 (34.4)	418 (65.6)	
	5 Databases	124 (2.4)	48 (38.7)	76 (61.3)	
	6 Databases	9 (0.2)	3 (33.3)	6 (66.7)	
Year of arrest	2001	480 (9.35)	111 (23.1)	369 (76.9)	$\chi2 = 22.807$, df=5, p-value = 0.0003674

	2002	866 (16.9)	152 (17.6)	714 (82.4)
	2003	1083 (21.1)	182 (16.8)	901 (83.2)
	2004	1244 (24.2)	173 (13.9)	1071 (86.1)
	2005	1186 (23.1)	202 (17.0)	984 (83.0)
	2006	274 (5.3)	54 (19.7)	220 (80.3)

For numerical variables, as shown by **Table 2** below, the median *age* of the people arrested was 22, the youngest person was 13, and the oldest was 67. The standard deviation of age was 8.33.

Table 2: Descriptive statistics for held numerical predictors. The p-value shown is for the likelihood ratio test against the null intercept model.

Predictor	Median	Stdev	Min	Max
Age of person arrested for possession of small amount of marijuana	22	8.330865	13	67

4. Results

4.1 Two-Way Contingency Analysis

As shown in **Table 1**, we conducted two-way contingency analyses and chi-squared tests of association for each of the potential predictor variables with the outcome variable, given that each cell has sufficient expected counts. Both from examining the contingency tables and from the p-value derived from the association tests, it is obvious that with a p-value less than 0.05, race, employment status, citizenship, count of databases, and year are not independent of the outcome variable, while sex, prior traffic convictions, and regional districts seem to be less associated with the outcome variable.

4.2 Mosaic Plots

We continued the analysis by drawing mosaic plots for each categorical variable with the outcome variable.

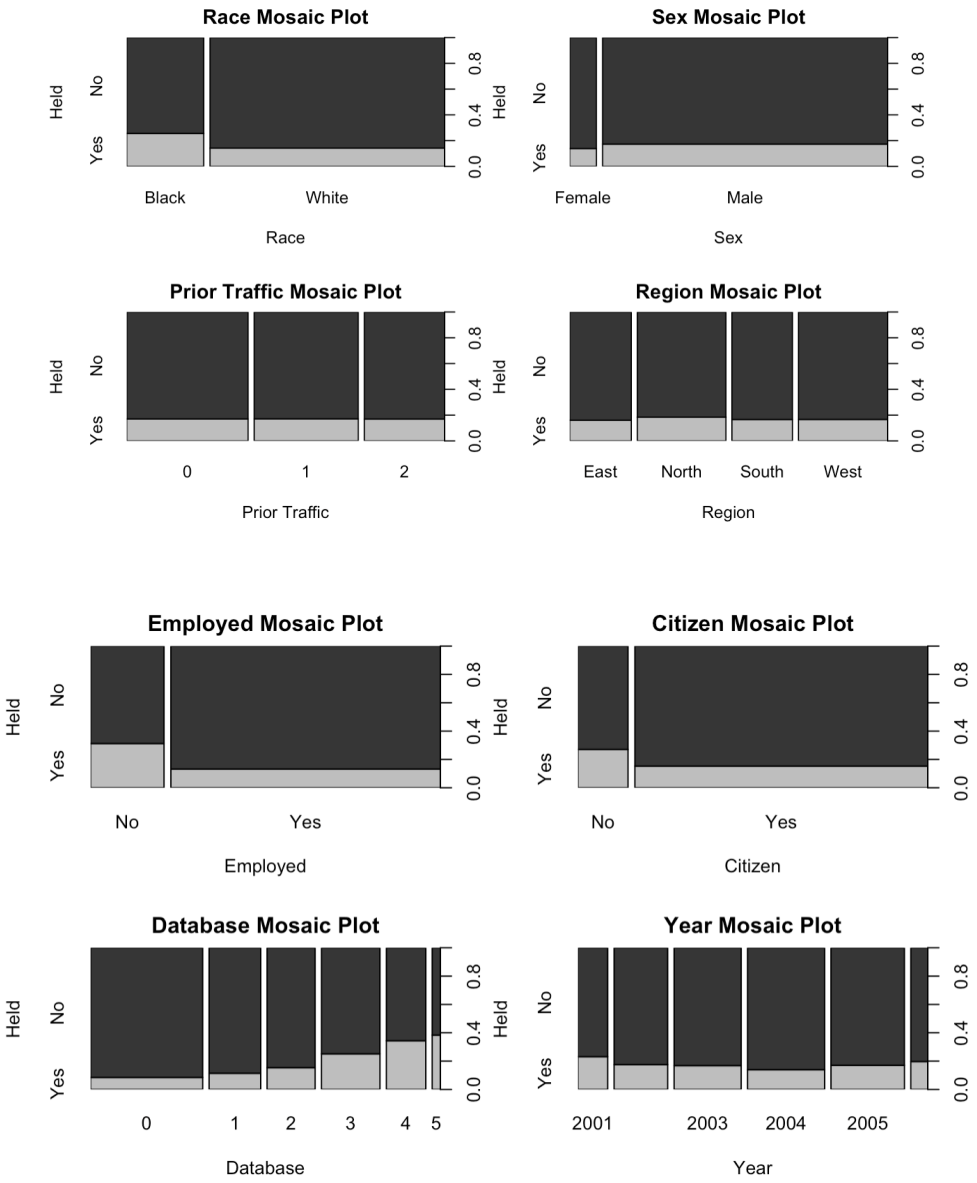


Figure 1: Mosaic plots of categorical variables

From the mosaic plots, it appears that *race*, *employed*, *citizen*, *year*, and *databases* lead to significant differences in the probability of being held for worse punishments. *Databases* exhibit a linear trend: having appeared in more police databases tend to have a higher probability of being held in jail or in the police station. *Sex*, *prior.traffi*c, *region* have seemingly less distinguishable differences in the probability of being held for worse punishments. The mosaic plots confirm the results of the two-way contingency analysis from the previous section.

4.3 Transformation of Numerical Variables

There is only one numerical variable: *age*. Thus, we prepared a "slicing-dicing" plot of the log-odds of being held for a worse punishment, shown below.

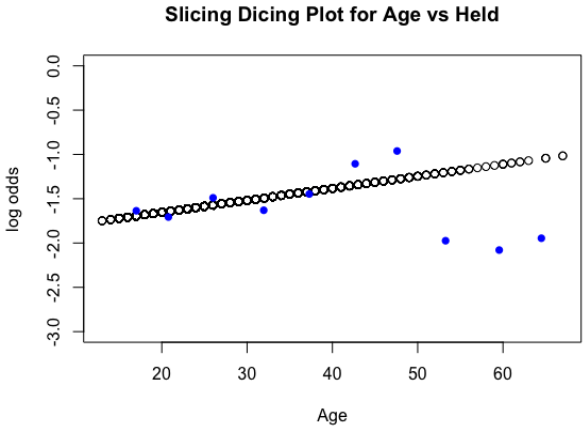


Figure 2: Slicing-dicing plot for age

The pattern is not linear, nor any other defined form either. Given this, we explored the plausibility of interactions and found that age and race appear to have an interaction effect. We drew a "slicing-dicing" plot for *age*, separately for white and black people. As shown by the plots, for white and black people separately, the log-odds of being held for a worse punishment have an approximately linear trend. Hence, there is no need to conduct a transformation for the *age* variable when we include the interaction effect between *age* and the *outcome* variable.

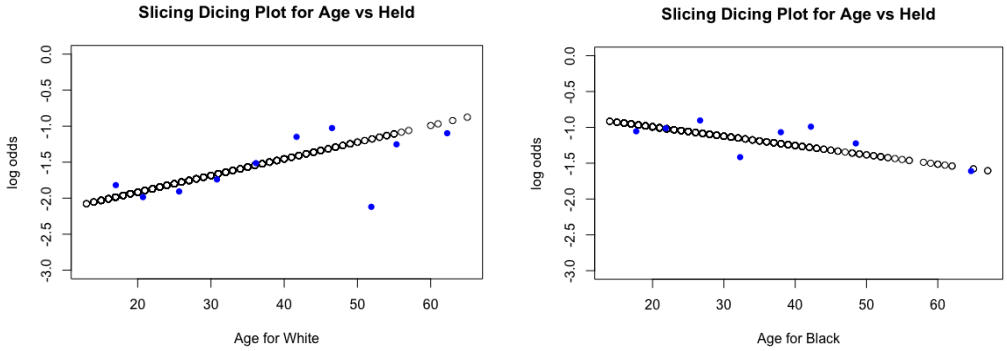


Figure 3: Slicing-dicing plot for age separated by race

4.4 Combined Levels of Categorical Variables

After cleaning and exploring each of the explanatory variables, we discovered that the *databases* variable has only 9 data points when *databases* is equal to 6. To ensure balanced levels with sufficient data, we combined level 5 and level 6. With this combination, the *databases* variable has a total of 5 levels ranging from 1 to 5.

4.5 Model Selection and Multivariable Analysis

We started building the model by first adding *race*, *age*, and the interaction between the two, as shown by the contingency analysis and the "slicing-dicing" plot described above. Then, we conducted forward selection based on the initial model. The decision criterion is that we only keep the variable once at a step with the lowest p-value of the likelihood-ratio test. Finally, we used AIC to compare the best model selected with different numbers of variables. As shown by the contingency analysis, we limited the search for potential variables within the significant ones, namely, *race*, *age, citizen, employed, databases*, and *year*. The first variable added to the base model is *databases* (with dummy coding). Then, the next variable added is *employed*. Notably, *employed* has an interaction effect with the *age* variable, as shown in the slicing-dicing plot below.

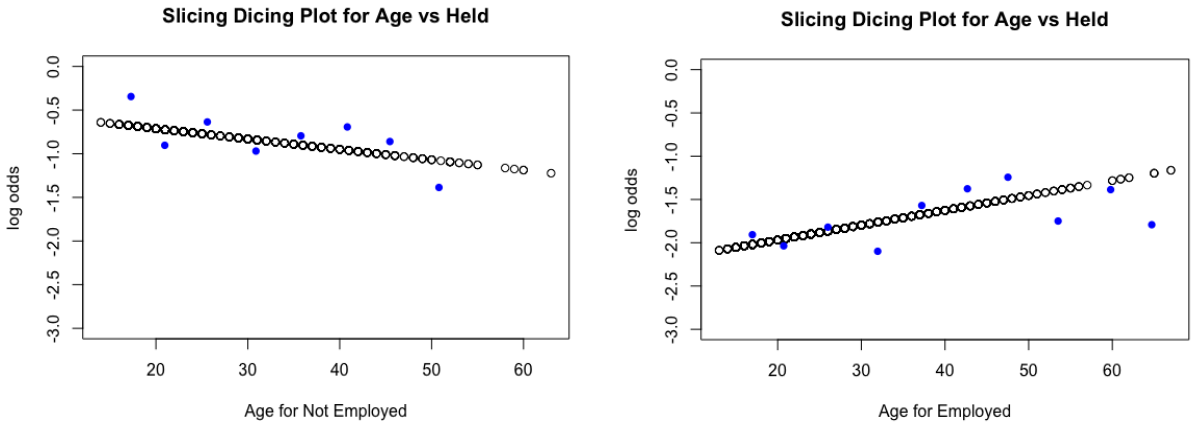


Figure 4: Slicing-dicing plot for age separated by employment status

Then, we added the *citizen* variable. Finally, we found that the p-value of the likelihood-ratio test for *year* is not significant, so we did not add it to the model. The AIC for the model decreases when we add each of the variables, so we keep the final model. The models considered during the final model selection and the related criterion and process are summarized in the following table:

Table 3: Models considered during p-values are obtained from the likelihood ratio test. Each model adds one new model at a time. A significant (green) variable is added to the model. If not significant (red), the variable is not added. * denotes $p < 0.05$, ** denotes $p < 0.01$, *** denotes $p < 0.001$.

#	Predictors	Added P-values	AIC	BIC	DF	Dev	AUC
1	age, race, age:race	<2e-16 ***	4588.8	4614.9	5129	4580.8	0.595
2	age, race, age:race, databases	<2e-16 ***	4331.6	4390.5	5124	4313.6	0.698
3	age, race, age:race, databases, employed	<2e-16 ***	4254.4	4319.8	5123	4234.4	0.717
4	age, race, age:race, databases, employed, citizen	1.292e-08 ***	4224.0	4296.0	5123	4234.4	0.725

5	age, race, age:race, databases, employed, citizen, year	0.278	4227.7	4332.4	5117	4195.7	0.728
6	age, race, age:race, databases, employed, citizen, age: employed	0.01589 *	4220.2	4298.7	5121	4196.2	0.726

The general form of the final model is as follows:

$$\text{logit}(\pi_d) = \beta_0 + \beta_W \text{race} + \beta_E \text{employed} + \beta_C \text{citizen} + \beta_{D1} \text{databases1} + \beta_{D2} \text{databases2} + \beta_{D3} \text{databases3} + \beta_{D4} \text{databases4} + \beta_{D5} \text{databases5} + \beta_A \text{age} + \beta_{WA} \text{race:age} + \beta_{EA} \text{employed:age}$$

where the continuous variable is:
age = person's age in years
and where the categorical variables were split into predictors using dummy coding as follows:
databases1-5 = indicating how many police databases the person appeared in
race = 1for white, 0 for black
employed = 1 for employed, 0 for unemployed
citizen = 1 for US citizen, 0 for non-US citizen
and where the interaction effects are as follows:
race:age = interaction effect between race and age
employed:age = interaction effect between employed and age

Then, a table is created with parameter estimates, odds ratios, odds ratio confidence intervals, p-value for each variable in the final model.

Table 4: Coefficient estimates, odds ratios, odds ratio 95% confidence intervals, and p-values for the final logistic model. Odds ratio confidence intervals are obtained from 95% Wald confidence intervals for the coefficients, p-values are obtained from the Wald test. * denotes $p < 0.05$, ** denotes $p < 0.01$, *** denotes $p < 0.001$.

Variable	Estimate	Odds Ratio	Odds Ratio 95% CI	p-value
Intercept (β_0)	0.11	1.11	(0.82,1.52)	0.725
race (β_W)	-1.25	0.29	(0.22,0.38)	04.74e-06 ***
employed (β_E)	-1.36	0.26	(0.2,0.33)	4.13e-07 ***
citizen (β_C)	-0.59	0.55	(0.5,0.61)	5.95e-09 ***
databases1 (β_{D1})	0.23	1.26	(1.1,1.45)	0.09828 .
databases2 (β_{D2})	0.5	1.66	(1.45,1.89)	1.67e-04 ***
databases3 (β_{D3})	1.06	2.87	(2.55,3.23)	<2e-16 ***
databases4 (β_{D4})	1.46	4.29	(3.79,4.86)	<2e-16 ***
databases5 (β_{D5})	1.60	4.96	(4.05,6.09)	5.00e-15 ***

age (β_A)	-0.04	0.96	(0.95,0.97)	1.29e-04 ***
race:age (β_{WA})	0.03	1.03	(1.02,1.04)	9.46e-04 ***
employed:age (β_{EA})	0.02	1.02	(1.01,1.03)	1.63e-02 *

Finally, we assessed the overall performance of the model using classification table, goodness-of-fit test, and ROC curve. The classification table is as follows:

Table 5: Classification table

	True	
Predicted	No	Yes
No	4207	807
Yes	52	67

From the data we have 874 successes, which means that the model we have adheres to the guidelines of having at least 10 successful observations per predictor. The classification table helps to determine the accuracy, specificity, and selectivity. The accuracy is (4207+67)/5133=0.833. The specificity is 4207/(4259)=0.988. The selectivity is 807/(807+67) = 0.923. These numbers indicate that the final model is sufficiently reliable since it does not produce too many false positives or false negatives.

The ROC curve is as follows

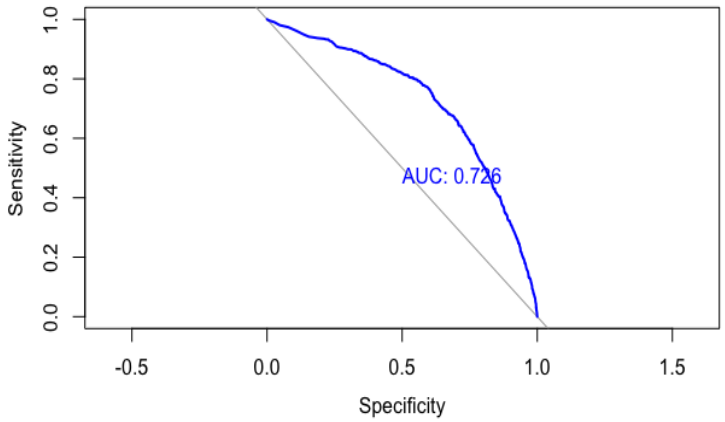


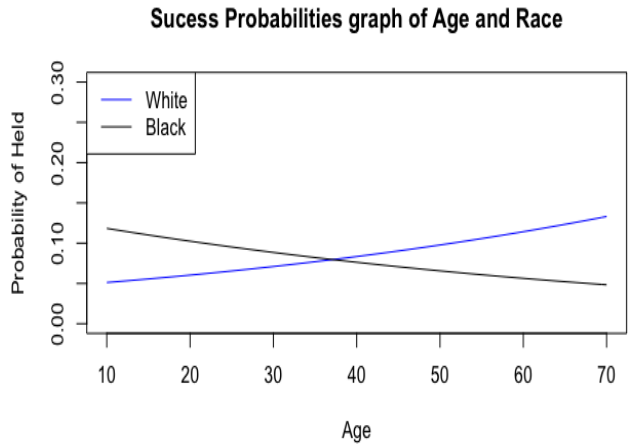
Figure 5: ROC plot, where AUC is 0.726

The AUC of the ROC curve is 0.726, which is higher than the diagonal line, showing that the performance of the final model is much better than the random guessing.

Finally, we also conduct the goodness-of-fit test to further assess the overall performance of our final model. The residual deviance is 4196.2 on 5121 degrees of freedom. So the p-value (0.999) indicates that the final model has a very good performance.

4.6 Sub-population Analysis

In this section, we presented three predicted probability plots: one for *employed* and *age*, one for *race* and *age*, and the other for *citizen* and *age*. In each of these plots, the x-axis is the value of *age* variable, the y-axis shows the probability of being held for more severe punishments, and each line in the graph corresponds to each level of *race*, *employed*, and *citizen*, respectively. Variables not included in the plots are kept constant at the mode.



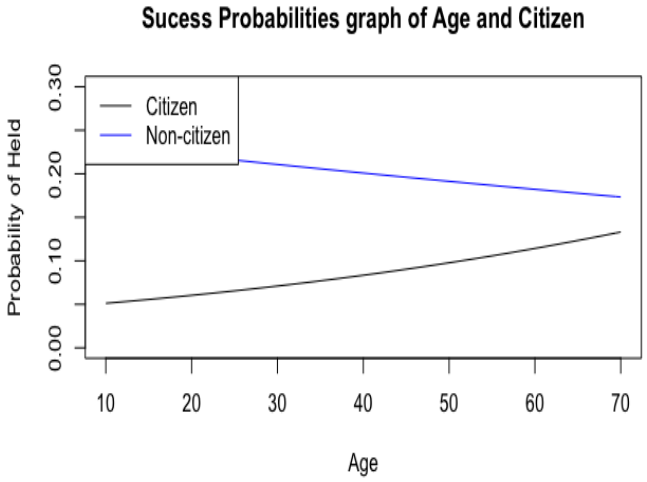


Figure 6: Probability plots of representative sub-population: race (upper), employed (middle), and citizen (lower)

5. Discussion

5.1 Effect and Direction of Variables

Holding other variables constant, being employed is associated with a decrease in the odds of receiving a severe treatment by a factor of exp(-1.36+0.02*age). Specifically, for a person of 20 years old, holding other variables constant, being employed is associated with a decrease in the odds of receiving a severe treatment by a factor of 0.38.

Holding other variables constant, being white is associated with a decrease in the odds of receiving a severe treatment by a factor of exp(-1.25+0.03*age). Specifically, for a person of 20 years old, holding other variables constant, being employed is associated with a decrease in the odds of receiving a severe treatment by a factor of 0.52.

Holding other factors constant, being a US citizen is associated with a decrease in the odds of being held for more severe punishments by a factor of 0.55.

Holding other factors constant, compared to not being in any police database, being in 1 police database is associated with an increase in the odds of being held for more severe punishments by a factor of 1.26. Holding other factors constant, compared to not being in any police database, being in 2 police databases is associated with an increase in the odds of being held for more severe punishments by a factor of 1.66. Holding other factors constant, compared to not being in any police database, being in 3 police databases is associated with an increase in the odds of being held for more severe punishments by a factor of 2.87. Holding other factors constant, compared to not being in any police database, being in 4 police databases is associated with an increase in the odds of being held for more severe punishments by a factor of 4.29. Holding other factors constant, compared to not being in any police database, being in 5 police databases is associated with an increase in the odds of being held for more severe punishments by a factor of 4.96.

Holding other factors constant, for an unemployed black person, being one year older is associated with a decrease in the odds of being held for more severe punishments by a factor of 0.96. Holding other factors constant, for an unemployed white person, being one year older is associated with a decrease in the odds of being held for more severe punishments by a factor of 0.99. Holding other factors constant, for an employed black person, being one year older is associated with a decrease in the odds of being held for more severe punishments by a factor of 0.98. Holding other factors constant, for an employed white person, being one year older is associated with a decrease in the odds of being held for more severe punishments by a factor of 1.01.

5.2 Racial Profiling

Based on the aforementioned analysis, racial profiling is evident. First, based on the contingency analysis and mosaic plot, African American people have a significantly higher probability of being held for more severe punishment. Furthermore, based on the final model we built, the main effect and the interaction effect of the *race* variable are significant. Being black correlates with a higher probability of being held for more severe punishments, especially for young people, holding other factors constant. Specifically, for a person of age 20, being white lowers the odds of being held for a higher punishment by a factor of 0.52 (almost half) while holding other factors constant.

5.3 Problems Encountered

On one hand, it can be hard to select the predictors that make the model effective yet parsimonious at the same time. Different criteria sometimes indicated different sets of variables, so the decision on selecting the final model relies on a holistic evaluation of the model performances. On the other hand, it is critical to eliminate bias when analyzing data collected from social issues. Some assumptions and stereotypes we had could be misleading. It is crucial to focus on the evidence provided by the data themselves.

5.4 Further Research

This study demonstrates a basic framework for analyzing differential treatments due to arrests for marijuana possession, providing many paths for future study. First, data collected from 2001 to 2006 can be constrained due to the limited time span. Utilizing data collected over a longer time frame may lead to interesting findings. Furthermore, since the culture can be drastically different in different areas of the United States, it can be instructive to conduct panel data analysis using data from multiple regions to gauge the phenomenon. Lastly, a complicated social issue such as the one analyzed in this study may involve other confounding variables that can distort the results. Thus, based on the framework built in this study, further research can be conducted using data that reflect more aspects of the situation.