Regression and Contingency Analysis in Arrest for Marijuana Possession Dataset

Minghan Liang, Kris Zhang, Yuze(Derek) Gu

Project Advisor: Melissa Smith

Abstract

When marijuana was not legalized, what determined whether an arrestee would be released or be held in jail was basically the police's discretion. The question is what factors are predictive of influencing a police's discretion and hence determining the consequences of an arrestee.

The data we investigate consists of 5140 individuals with various demographic and socioeconomic features who were arrested for possession of marijuana from 2001 to 2006. We adopt a logistic model to fit the data.

We conclude that the factors that are predictive of an arrestee's treatment include race, region, databases, citizen, employed, age, and the interaction term of race and region based on our final fitted model. Additionally, we investigate whether racial profiling impacts the treatment decision. We conclude that there is statistically significant evidence for two clear trends of racial profiling. The first trend is that males are more likely to experience racial profiling than females, controlling for employment status. The second one is that adults are more likely to encounter racial profiling than juveniles, while previous conviction record further increase the likelihood.

Introduction

Although marijuana has been legalized across many US states in recent years, possessing even a small amount of marijuana could lead to legal consequences in the early 2000s. Depending on the scenarios and backgrounds of arrestees, the police could either choose to release the arrestee with a summon to appear in court or implement harsher treatment, such as detention in the police station or held in jail. It is the police's discretionary power to choose whether to implement harsher enforcement. However, such a power can be potentially abused towards people from certain ethnic or underprivileged groups. Historically, the incarceration and arrest rates are nearly 5 times higher for African Americans than white Americans due to various complex socioeconomic issues such as unemployment, previous misdemeanors, and poor education.

In this study, our primary goal consists of two analytical objectives: First and most importantly, we aim to construct a classification model using the Logistic Regression model architecture where we predict whether a marijuana-processed arrestee will receive harsher law enforcement (such as

detention and prison held), based on information such as race, age, sex, employment status, and previous conviction records. Secondly, we aim to analyze whether racial profiling exists in police discretionary action based on our fitted regression model and an additional test of association between harsh law enforcement and arrestees' race, controlling for other factors. The analytical tasks in this project will be performed on a dataset that records two different types of law enforcement consequences (being held or not held) concerning the arrestees' socioeconomic and personal information collected from 2001 to 2006.

Data Exploration

Before entering the analysis part, the data is checked and cleaned. The main objective for this part is to investigate the distribution of our features and check whether any problematic data point might interfere with the analysis and disrupt the model. In the actual process of checking, we find that there are a few data points that are irregular or irrational. For instance, there is an individual labeled with the race "Gr" where there are supposed to be only two responses (Black and White) for the race predictor. Other suspicious instances include an individual with the region "Purple", one with age 117, and another with a number of police databases of 33. Since these irregular data points can potentially lead to a poor fit, we take these sets out of the data to improve the integrity of our analysis.

There are two numerical predictors: databases and age. Databases range from 0 to 6 with a median of 1. Most individuals tend to appear in a few police databases. Age ranges from 13 to 67 with a median of 22. The age distribution has a right tail since more younger people are arrested than older people. For numerical variables, we also investigated the linearity of their values with respect to the held ratio. As a result, there exist clear linearity between held ratio and age & databases since the slicing graph shows a clear upward linear trend (See Figure#2)

Predictors	databases	age
Min	0	13
Max	6	67
Median	1	22
Standard Dev	1.5402	8.3309

Figure 1: Summary of Variables, Numerical

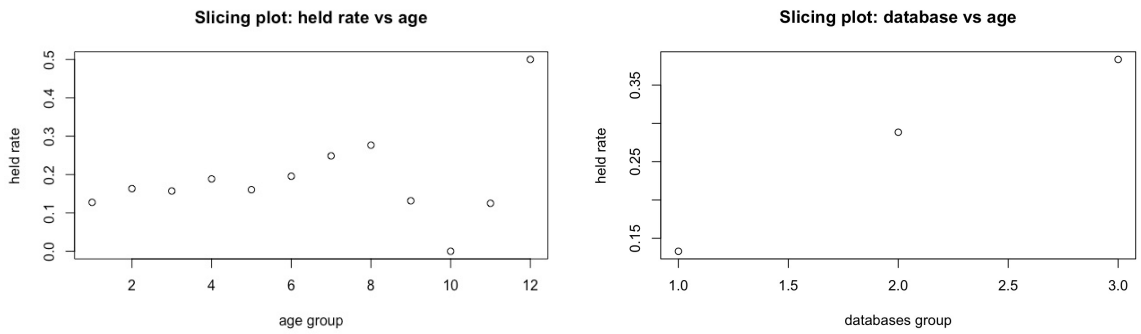

Figure 2: Summary of Variables, Numerical

There are six categorical predictors in total. For the race predictor, the number of white individuals is 3.05 times that of black individuals. The proportion of black individuals who were held is 11.4% larger than that of white individuals. For the sex predictor, the number of males is much larger (10.69 times) than that of females. The proportion of males who were held is slightly (3.42%) larger than that of females. For the region predictor, the number of individuals in each of the four categories is about even. The percentages of individuals held are also very close, where the north has the largest proportion held with 18.39% and the other three regions have almost identical proportions. For the employed predictor, the number of individuals employed is 3.67 times the number of individuals unemployed. The proportion of unemployed individuals held is 18.05% larger than that of employed individuals. For the citizen predictor, the number of U.S. citizens in the data set is 5.85 times the number of non-U.S. citizens. The proportion of non-U.S. citizens held is 11.79% higher than the proportion of U.S. citizens being held. Last but not least, the prior.traffic predictor contains three categories: no prior traffic conviction, one prior traffic conviction, and two or more traffic convictions. Of the three categories, the number of individuals with zero traffic conviction individuals is the largest, following the number of individuals with one traffic conviction, then the number of two or more convictions individuals being the smallest. The proportions of individuals held among the three groups are very close around 17%. One problem that appears in the data exploration process is that the numbers of individuals across various categories are not even for some predictors such as race, employment, and citizenship. This is one limitation of the dataset and can lead to potential biases in the model and analysis.

Predictor	Categories	Total	Held	Not Held
race	black	1269	325(25.61%)	944(74.39%)
	white	3864	549(14.21%)	3315((85.79%)
sex	male	4694	813(17.32%)	3881(82.68%)
	female	439	61(13.90%)	378(86.10%)
region	north	1517	279(18.39%)	1238(81.61%)

	south	1036	172(16.60%)	864(83.40%)
	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%)
	no	1099	343(31.21%)	756(68.79%)
citizen	yes	4384	671(15.31%)	3713(84.69%)
	no	749	203(27.10%)	546(72.90%)
prior.traffic (ordinal)	0	2038	347(17.03%)	1691(82.97%)
	1	1747	299(17.12%)	1448(82.88%)
	2 or more	1348	228(16.91%)	1120(83.09%)

Figure 3: Summary of Variables, Categorical

Models

The first goal of this project is to build a model that predicts the probability of held based on the nine predictors that were given in the dataset. As logistics regression was conducted on the dataset, the result is displayed in the Table below with its P-value. In this project, we use the threshold of 0.05 to determine the level of significance of our predictors. Among all nine predictors, four appears to be highly significant in predicting the outcome variable held. During model selection process, backward stepwise model selection was performed during the model selection stage, during which employment status, citizenship, and database were selected as the four significant predicting variables. Forward stepwise selection was also performed to determine predictors that could be added. During our selection process, we found that age alone is not a significant variable; however, when age and the interaction term on age and race was included, both of them become significant in our model. Besides of All other variables added does not seem to be significant. Below are models considered during the final model selection step. (See Figure)

- Predictor abbreviates: Rc(race), E(employment status), C(citizenship), D(databases), Y(year), S(sex), P(prior traffic), Rg(region), A(age)
- During Forward selection, we add a new variable to the previous model at each step, and we calculate the p-value of the newly added variable to reflect their statistical significance level. If it's significant (<0.05), it is labeled red; otherwise, it is labeled green. If the added p-value does not appear significant, it is not included in the next model. The added p value includes all predictors caused by the possible dummy coding for the added variable.

Predictors	AIC	P-value, Added predictor(s)	BIC	DF	Dev	Goodness of Fit	ROC
Rc+E+C+D	4231.7	—	4264.4	5129	4221.7	1	0.722
Rc+E+C+D+A	4231.7	0.721	4272.8	5128	4221.6	1	0.724
Rc+E+C+D+A+Rc*A	4225.2	0.002, <0.001	4276.5	5126	4208.2	1	0.725
Rc+E+C+D+S+A+Rc*A	4224.8	0.876	4277.1	5126	4221.7	1	0.725
Rc+E+C+D+P+A+Rc*A	4224.7	0.806	4277.1	5126	4208.7	1	0.725
Rc+E+C+D+Rg+A+Rc*A	4226.7	0.177, 0.634, 0.687	4277.1	5124	4206.7	1	0.725
Rc+E+C+D+Y+A+Rc*A	4224.2	0.749	4276.5	5126	4208.2	1	0.725

Figure 4: Summary of Models

After a general navigation on different models generated in the previous exploration, we started to construct our final model by discarding features that are not significant across all previously investigated models, and apply further feature engineering steps on some other important variables. First, we adjusted the prior traffic variable by changing all numbers that are greater or After general navigation on different models generated in the previous exploration, we constructed our final model by discarding features that are not significant across all previously investigated models and applying further feature engineering steps on some other important variables. First, we adjusted the prior traffic variable by changing all numbers greater or equal to 1 as 1 (there are records of prior traffic) and 0 otherwise (there are no records of prior traffic). However, it still does not appear to be a significant variable.

Second, we found that when the interaction between race and age was added, it appeared that the interaction between race and age was highly significant. Additionally, the AIC term appeared to be lower than the model without the interaction term. However, without the interaction, age is not a significant variable alone. Therefore, it could be because one unit change of this variable is not significant enough to make a difference in the held decision. To overcome this problem, we applied the binning method to the age variable. It was first grouped into 3 groups: less than 18 years old, between 18 and 40 years old, and greater than 40 years old. It appeared that only the p-value of fewer than 18 years old is less than 0.05. With this conclusion, we further optimized our method by grouping age variables into two groups: less than 18 years old and greater than or equal to 18 years old. As a result, the age variable is tested significant through the T-test in predicting the decision of held under this grouping (See table).

We also want to know if location influences the police's decision to hold one who possesses marijuana. However, the region does not appear to be a significant variable in our original model. Therefore, the variable is recombined into two groups: the North region(encoded as 1) and all others(encoded as 0). As a result, the interaction between race and region appears to be significant.

Predictor variables: A(adj): age with grouping into less than 18 years old(underage) and greater than or equal to 18 years old(adult). Rg(adj): region with grouping into North region(encoded as 1) and all others(encoded as 0)

Predictors	AIC	Added P value	BIC	DF	Dev	P-Value	ROC AUC
Rc+E+C+D	4231.7	—	4264.4	5129	4221.7	1	0.722
Rc+E+C+D+A(adj)	4222.8	<0.001	4260.7	5128	4209.4	1	0.726
Rc+E+C+D+A(adj)+Rg(adj)+ Rc*Rg(adj)	4219.7	0.016, 0.049	4272.0	5126	4203.7	1	0.726

Figure 5: Summary of Models, after calibration

The model with lower BIC have a slightly lower AIC. All models appear to fit very well according to the goodness-of-fit test and have very similar AUC. As a result, the model with lowest AIC and BIC was chosen as our final model.

Results & Findings

Part I. Regression Analysis

After careful examinations and selections, we constructed a best-performing final model, including race, employment status, citizenship, database, age, region, and race*region as model parameters. Below is the statistical expression of our final model and a summary of our predictors.

- **Race**: race of the person possessing marijuana. 1 for White, 0 for Black
- **Region**: region on the person found. 1 for region North, 0 for otherwise.
- **Databases**: count of 0 – 6, indicating how many police databases the person appeared in
- **Citizen**: 1 for citizen, 0 for non-citizen
- **Employed**: 1 for employed, 0 for unemployed
- **Age**: 1 for less than 18 years old(underage), 0 for greater than or equal to 18 years old.
- **race*region**: interaction term. 1 for both white and presenting in north region, 0 otherwise.

$$\log\left[\frac{P(\text{held}=1)}{1-P(\text{held}=1)}\right] = \alpha+\beta_1(\text{race})+\beta_2(\text{region})+\beta_3(\text{databases})+\beta_4(\text{citizen})+\beta_5(\text{employed})+\beta_6(\text{age})+\beta_7(\text{race}\times\text{region})$$

Predictors	Estimates	Odds Ratio	Odds Ratio CI	P-Value
(Intercept)	-1.18	0.31	0.23-0.40	<0.001
Race	-0.29	0.75	0.61-0.92	0.006
Region	0.35	1.42	1.07-1.88	0.016
Databases	0.37	1.45	1.38-1.53	<0.001
Citizen	-0.55	0.58	0.47-0.71	<0.001
Employed	-0.78	0.46	0.39-0.54	<0.001
Age	0.44	1.55	1.22-1.96	<0.001
Race*Region	-0.35	0.70	0.50-1.00	0.049

Figure 6: Summary of Models, after calibration

By the One-in-Ten rule, an ideal model can have a maximum of 87 possible predictors since our dataset consists of 874 positive outcomes, and our final model only consists of 7 predictors. To assess how well the model fits, We performed a Goodness-of-Fit test on the fitted model. As a result, we found that the residual deviance is 4203.7 on 5126 degrees of freedom, and the p-value is 1(>0.050), which indicates a great fit of the model.

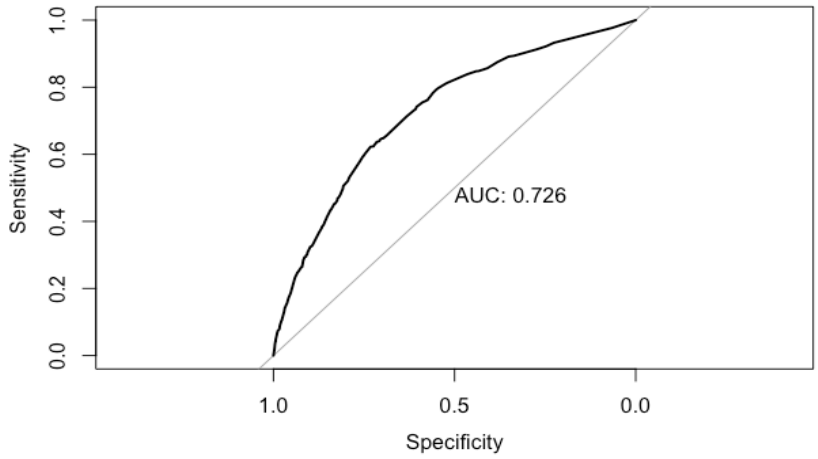

Figure 7: Summary of Models, after calibration

The classification table (see figure) with a default prediction threshold (0.5) shows that our model's specificity is 0.56 and sensitivity is 0.84. While we want both of them to be close to 1, our model achieves a fairly good result with a false positive rate of 0.16 and a false negative rate of 0.44. Additionally, the Area Under the Curve (AUC) appeared to be 0.726, which also indicates a good fit of our model.

Figure 8: Classification Results, Final Model

	Actual	
Predicted	**No**	**Yes**
No	4209	51
Yes	810	64

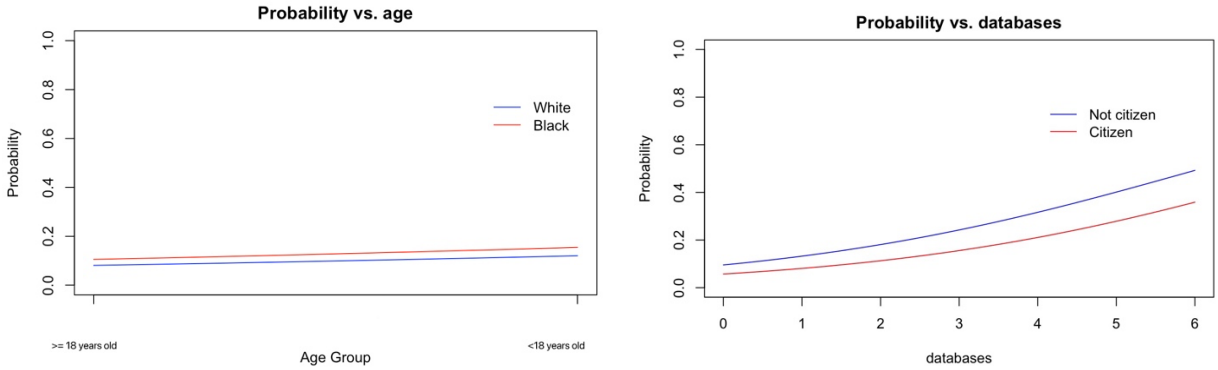

Figure 9: Probability Curves, Held Probability vs Age (left) and Database(Right)

Part II. Analysis on Racial Profiling

Overall, We found that there is a clear trend of racial profiling in law enforcement on marijuana-processed arrestees, proven by multiple statistical evidence. In conclusion, we found that the odds of receiving harsher law enforcement for African Americans is higher than for White Americans by a multiplicative factor of 1.33 (see Figure #6) based on our best-performing logistic regression model, while all other factors contributed to the policies' discretionary judgments are fixed and equal. Additionally, our contingency analysis on all data records also indicated highly statistically significant evidence for the presence of racial profiling (p-value < 2.2e-16) out in the wild.

However, we would like to analyze further the characteristic of racial profiling in the case of harsh law enforcement on marijuana possession. Namely, we have conducted further contingency analysis to assess the effect of racial profiling among people of different backgrounds, and we surprisingly found that racial profiling is most prevailing in the following scenarios:

- **Male is much more likely to encounter racial profiling than female, controlling for employment status.** We believe that racial profiling is highly likely to occur among males due to males' traditional narrative in western culture and the historical problem of gender bias in the U.S. Moreover, such an injustice would be more likely to occur among unemployed individuals due to a systematic bias toward the association between unemployment and drug dealing in today's U.S. society. As a result, we found our statistical evidence indicates that racial profiling is highly significant among males, both employed and unemployed (See Figure #10)

No	Gender	Employment Status	Chi-Squared Test statistic (Degree of Freedom=1)	P-Value
1	Male	Yes	44.253	<0.0001
2	Male	No	18.379	<0.0001
3	Female	Yes	0.56382	0.4527
4	Female	No	0.08907	0.7654

Figure 10: Contingency Test, Gender & Employment factors

- **Adult is much more likely to encounter racial profiling than juveniles, while previous conviction record further increases the likelihood.** Besides gender and employment status, we also believe that racial profiling is highly likely to associate with age group since juvenile arrestees are more protected and tolerated under current U.S. judicial systems regardless of their race. However, any previous conviction records (indicated by occurrences of personal information in one or more other police databases) would severely undermine the fairness of law enforcement, regardless of race and age. As a result, our statistical evidence indicates that racial profiling is only significant among adult groups without previous conviction records (p-value is 0.03035 for arrestees aged 19-22 and 0.00091 for arrestees aged 23 or more). On the other hand, racial profiling is significant among all age group if arrestees possess any previous conviction records (See Figure #11)

No	Age Group	Previous Conviction (databases >= 1)	Chi-Squared Test statistic (Degree of Freedom=1)	P-Value
1	<=18	No	0.486	0.48570
2	<=18	Yes	11.950	0.00055
3	19 - 22	No	4.695	0.03025
4	19 - 22	Yes	12.987	0.00031
5	>= 22	No	11.002	0.00091
6	>= 22	Yes	14.736	0.00012

Figure 11: Contingency Test, Age & Previous Conviction factors

Future Works

In this project, we successfully constructed a classification model on the marijuana possession dataset with the Logistic Regression Model Architecture and discovered statistical evidence that reveals important trends in racial profiling through the dataset. However, this project has many potential opportunities to improve the data analysis process and classification model.

First and most importantly, there are many potential opportunities to improve the feature qualities in our given dataset by introducing additional feature engineering procedures. In this project, we transformed several numerical and categorical features through binning and level combination based on feature importance during the model fitting stage. However, we did not leverage all the domain knowledge associated with features in the current dataset, as some are either not fully interpreted or hard to quantify. For instance, we found that prior traffic and year information failed to exhibit statistical significance in any of the fitted models throughout the project. We believe that those features are worth to be further investigated if we could associate them with proper exterior knowledge (such as related laws or promulgation of new laws related to possession of marijuana during 2001-2006 within the region represented by our current dataset) and transform them in proper form.

Secondly, we can expand our contingency analysis framework by introducing additional analysis on additional features. Due to the limit of computational capacity, we only investigated the occurrence of racial profiling with respect to a few important factors, such as gender, employment, and age. However, it is worth further investigating other factors in our dataset to understand the presence of racial filing better. For instance, it is also worth investigating whether racial profiling is particularly prominent in particular regions or years since law regulations vary significantly across different regions and times.

Last but not least, it would be valuable to expand our analysis framework by collecting additional data and introducing new features to the dataset for better prediction results. In our current dataset, the majority of features only describe biographical information on the arrestees. However, other information such as educational level (less than high school, high school, college, etc.), crime scene (home, car, bar, etc.), and previous illegal drug possession history (Yes, No) would also be highly informative to be considered in our classification task. Should opportunities come to collect those data, our existing model framework would have a potential increase in prediction accuracy.