Model Selection & Cross-Validation for Marijuana Arrest Data

1.

Abstract

Brief overview of the project

Abstract

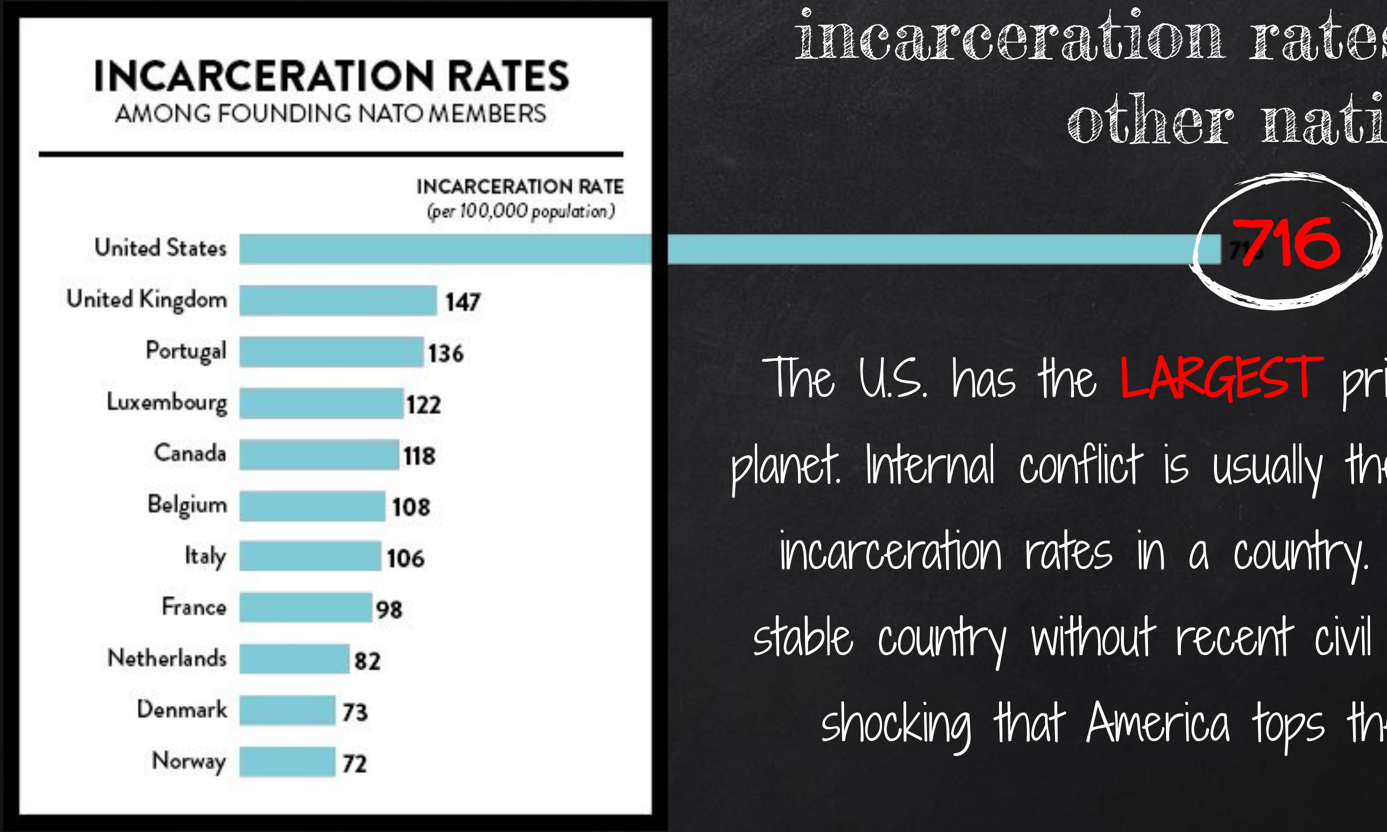
As the topic of marijuana use continues to gain attention on a global scale, the ramifications of such use consequently gains both social and scientific importance. The consequences of criminal arrest due to marijuana possession, particularly whether or not the arrestee was released with a summons, is discussed in the following presentation. To see which variables in our dataset predicted our outcome best, logistic forward, backward, and bi-directional model selection as well as cross-validation measures were implemented with the use of R programming. From strictly looking at the three models, it was very difficult to determine which independent variables best predicted the dependent variable since the results for each was the same. However, through cross-validation analysis, forward modeling was shown to be most effective due to higher kappa value.

2.

Introduction

Dataset Overview - Social Importance - Scientific Relevance

How do the United States' incarceration rates compare to other nations?



The U.S. has the **LARGEST** prison population on the planet. Internal conflict is usually the best predictor of high incarceration rates in a country. Thus, as a politically stable country without recent civil wars, it is particularly shocking that America tops the incarceration list.

Sources: Kelly, M. (2019, June 26). *How many people are in prison on marijuana charges?*. The Washington Post. Retrieved April, 10, 2020, from https://www.washingtonpost.com/politics/2019/live-updates/general-election/fact-checking-the-first-democratic-debate/how-many-people-are-in-prison-on-marijuana-charges/?arc404=true

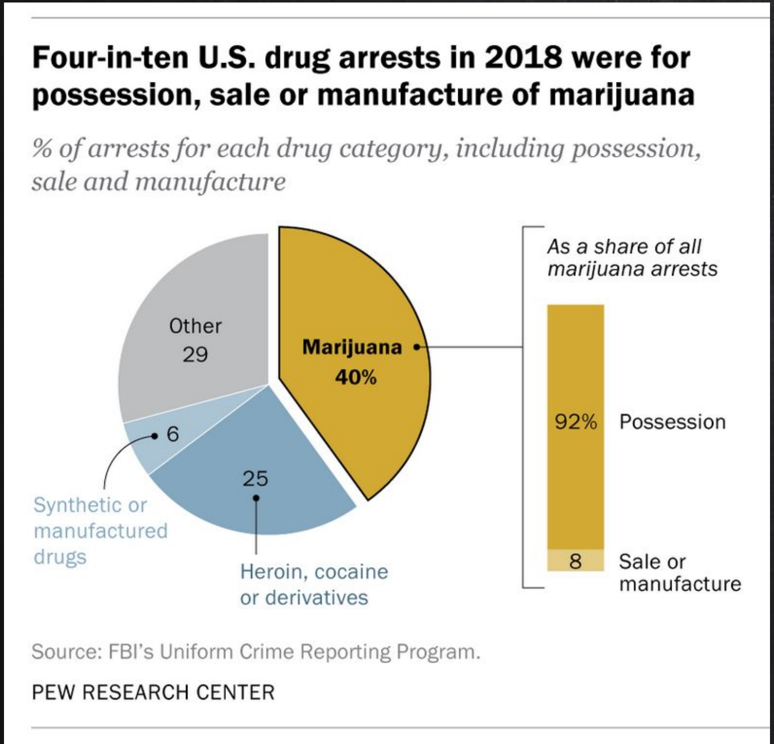
Wagner, P., Sakala, L., & Begley, J. (n.d.). *States of Incarceration: The Global Context*. Prison Policy Initiative. Retrieved April 10, 2020 from https://www.prisonpolicy.org/global/

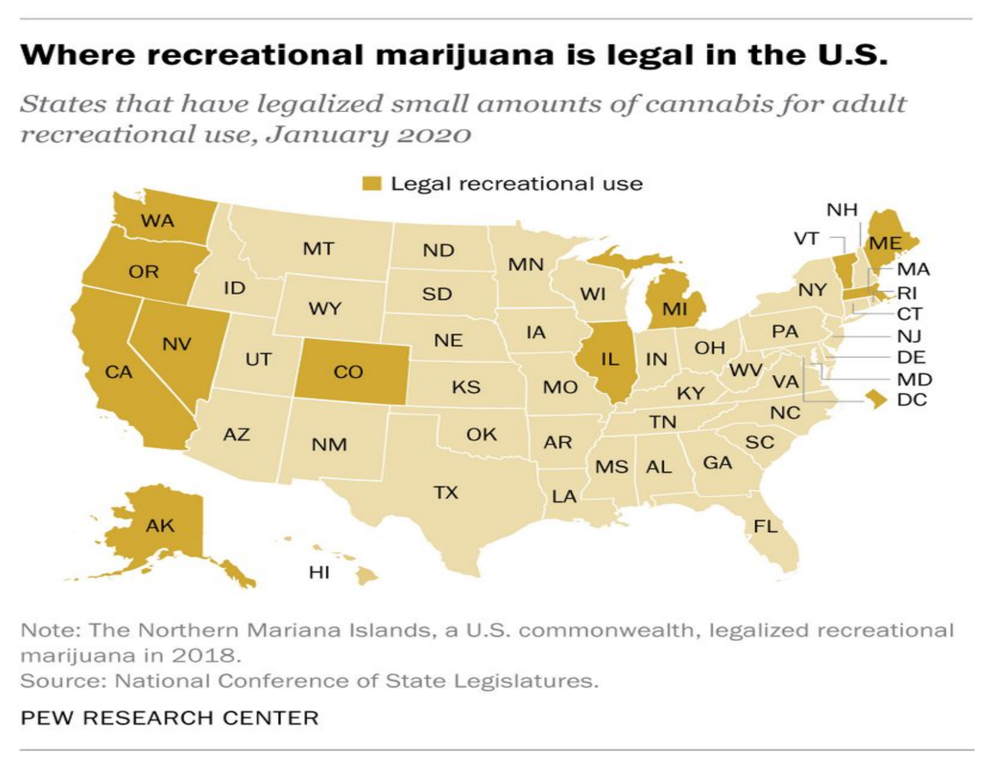
How do rates in NJ and NY compare to other nations?

New Jersey and New York's rates of incarceration are 506 and 492 out of every 100,000 population, respectively. These rates are similar to those of Cuba's, and both NJ and NY have higher rates of incarceration than Rwanda and the Russian Federation.

Source: Wagner, P., Sakala, L., & Begley, J. (n.d.). *States of Incarceration: The Global Context*. Prison Policy Initiative. Retrieved April 10, 2020 from https://www.prisonpolicy.org/global/

Marijuana Legalization + Drug Arrests, Nationally





Source: Gramlich, J. (2020). Four-in-ten U.S. drug arrests in 2018 were for marijuana offenses – mostly possession. *Pew Research Center*. Retrieved March 31, 2020, from https://www.pewresearch.org/fact-tank/2020/01/22/four-in-ten-u-s-drug-arrests-in-2018-were-for-marijuana-offenses-mostly-possession/

Even worse, drug arrests disproportionately affect people of color :

Despite making up just 31.5% of the U.S. population, the percentage of Black or Latino people arrested for drug law violations is 46.9%.

Source: Drug Policy Alliance. (n.d.). Drug War Statistics. Retrieved April 10, 2020 from https://www.drugpolicy.org/issues/drug-war-statistics

Marijuana: A Hot Topic in NJ

"Although I remain disappointed in the Legislature's inability to legislatively legalize adult-use marijuana, I am optimistic that the people of New Jersey, who overwhelmingly support legalization, will vote to do so. And, when they do, we will take a critical and long overdue step for real criminal justice reform."

New Jersey Marijuana

Legal weed is now up to N.J. voters as lawmakers vote to put it on 2020 ballot

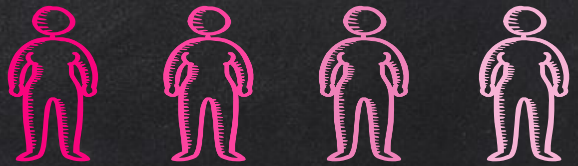
Updated Dec 16, 2019; Posted Dec 16, 2019

Gov. Phil Murphy made legalizing marijuana for those over 21 one of his campaign promises. In the nearly two years since he took office, the initiative has seen several setbacks. State Senate President Stephen Sweeney announced in late November he would not take the bill to the floor, and would instead seek to put it to the ballot for voters to decide.

Sources: Hoover, A. (2019, Dec 6). *Legal weed is now up to N.J. voters as lawmakers vote to put it on 2020 ballot.* NJ.com. Retrieved April 10, 2020 from https://www.nj.com/marijuana/2019/12/voters-could-decide-if-nj-will-legalize-weed-after-senate-votes-to-put-it-on-2020-ballot.html

The State of New Jersey. (2019, Nov 26). *Statement by Governor Murphy on Marijuana Decriminalization.* Official Site of the State of New Jersey. Retrieved April 10, 2020 from https://www.nj.gov/governor/news/news/562019/approved/20191126b.shtml

How does our project relate to our education?



Although the four of us come from a variety of majors and interests, we were all drawn to this topic because of its prominence in the national, state, and local news. As public health majors, Alex and Celine were interested in the social structures at play in America. Incarceration, addiction, and disparities each highlight topics often discussed in our public health classes. As BAIT and Math majors, Anna and Rachel were interested in the data science behind these issues. Moreover, as young adults at a majority liberal university in a 'blue' state, marijuana legalization is a focal point in many politicians' platforms. From the case built in the previous slides, it is clear to see that there is a problem with incarceration in the U.S., and especially incarceration due to drug arrests. We see the inequities, and we hope to promote awareness of the issue and to investigate it further through our project.

Overview of Chosen Statistical Topic

- We chose model selection for our project because we wanted to see which variables in our dataset modeled, or predicted, our outcome best, and which variables created the best fit. Cross-validation will help us to assess the accuracy and validity of our model; it will demonstrate which model is predicting best in comparison to the other models + will help us avoid over-fitting.
    - The principle of parsimony suggests that the model with the least variables but with the greatest explanatory power is the most useful
    - Additionally, in terms of real-world application, and experimenter preference, having less variables in your model can save time, money, and resources.

Dataset Description

- CHOSEN DATASET: Arrests for Marijuana Possession
  - Data on police treatment of individuals arrested in Toronto (Canada) for simple possession of small quantities of marijuana
  - From the carData package in R
  - Sample size: n = 5,226 observations
  - 8 variables total
    - dependent variable in ORANGE, independent variables in BLUE

Source: Friendly, M. (n.d.). Arrests for Marijuana Possession. [Dataset]. York University. http://math.furman.edu/~dcs/courses/math47/R/library/effects/html/Arrests.html

Variables in the Dataset

VARIABLE NAME	VARIABLE DESCRIPTION
released	Whether or not the arrestee was released with a summons; a factor with levels (yes, no)
color	The arrestee's race; a factor with levels (black, white)
year	1997 through 2002; a numeric vector
age	In years; a numeric vector
sex	A factor with levels (female, male)
employed	A factor with levels (yes, no)
citizen	A factor with levels (yes, no)
checks	Number of police databases (of previous arrests, previous convictions, parole status, etc. --- 6 in all) on which the arrestee's name appeared; numeric vector

Source: Friendly, M. (n.d.). Arrests for Marjiuana Possession. [Dataset]. York University. http://math.furman.edu/~dcs/courses/math47/R/library/effects/html/Arrests.html

Variables of Interest

From this data, we chose "released" as our dependent variable, and used the rest of the the variables as our potential predictors, or independent variables (i.e. colour, year, age, sex, employed, citizen, and checks).

For our project, we were interested in determining which characteristics (IVs) best predicted the release (DV) of an individual arrested in Toronto (Canada) for simple possession of small quantities of marijuana. Although our chosen dataset is not from the U.S., we thought it could help raise awareness about problems with incarceration worldwide.

3.

Materials + Methods

All about the code

Model Selection Explained

- Forward step-wise selection
  - This process selects ONE variable at at time to be added to the model
  - As long as the p-value of the parameter is less than 0.30, the parameter gets included in the model
- Backward step-wise selection
  - This process begins with all the parameters in the model and removes ONE parameter at a time
  - Parameters with p-values greater than 0.3 are removed until there are no candidates with a p-value above 0.3
- Bi-directional step-wise selection
  - This process removes and adds parameters at the same time with the above p-value specifications

Step 1: Loading Dataset Into R

> #Looking at the "Arrests for Marijuana Possession" dataset using R

> library(carData) #Companion to Applied Regression Data Sets

> library(epiDisplay)#to use the function logistic.display for an easier view of the logistic model

> library(StepReg)#package for model selection of a logistic regression model

> summary(Arrests) #summary of the "Arrests for Marijuana Possession" dataset

```
 released      colour          year           age            sex        employed   citizen       checks     
 No : 892   Black:1288   Min.   :1997   Min.   :12.00   Female: 443   No :1115   No : 771   Min.   :0.000  
 Yes:4334   White:3938   1st Qu.:1998   1st Qu.:18.00   Male  :4783   Yes:4111   Yes:4455   1st Qu.:0.000  
                         Median :2000   Median :21.00                                       Median :1.000  
                         Mean   :2000   Mean   :23.85                                       Mean   :1.636  
                         3rd Qu.:2001   3rd Qu.:27.00                                       3rd Qu.:3.000  
                         Max.   :2002   Max.   :66.00                                       Max.   :6.000  
```

> #There are 5226 observations with 8 variables

> data(Arrests, package="carData") #loads the specified data set

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 2: Modeling the Relationship Between Response + Predictors

> #We are trying to find a model that tells us the relationship of if someone who was arrested for marijuana possession was released with a summons and specific factors of the arrestee

> #We use a logistic regression since our dependent variable is binary categorical (Yes or No). Below is our complete model:

```
> logmodel <- glm(released~ colour+ year + age + sex + employed + citizen + checks, data = Arrests, family = "binomial")
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 3: Evaluate Our Complete Model

> anova(logmodel, test="Chisq")

> #By this point, we already have an idea that the most significant variables to y are checks, colour, employed, and citizen but we want to still want to see what models we will get if we do the different types of selection.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: released

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      5225     4776.3
colour    1   86.760      5224     4689.5 < 2.2e-16 ***
year      1    5.576      5223     4683.9   0.01821 *
age       1    5.886      5222     4678.0   0.01526 *
sex       1    1.375      5221     4676.7   0.24093
employed  1  152.255      5220     4524.4 < 2.2e-16 ***
citizen   1   22.527      5219     4501.9 2.072e-06 ***
checks    1  202.814      5218     4299.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 4: Can We Reduce The Amount of Independent Variables To Reduce Chance Of Overfitting?

> #We want to know if we need to include all independent variables of the dataset in this model.

> #Fewer variables means less variance and less variance means less chance of overfitting.

> #We will use model selection to find which independent variables are significant in predicting the dependent variable

> #Generate stepwise selection procedures on the 7 predictors provided in the logmod statement

Step 5: Forward Selection

> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "forward",
+ select = "AIC", sle = 0.15, sls = 0.15)

```
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1        checks                1        1 4461.5917
2    2      employed                1        2 4372.1808
3    3       citizen                1        3  4327.698
4    4        colour                1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
              Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
checks           checks  -0.3628     0.0257 -14.1008        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 6: Backward Elimination

```
> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "backward",
+                    select = "AIC", sle = 0.15, sls = 0.15)
```

```
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1                         sex  1        6 4313.0678
2    2                        year  1        5 4311.0896
3    3                         age  1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
              Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
checks           checks  -0.3628     0.0257 -14.1008        0
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 7: Bidirectional Selection

```
> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "bidirection",
+                    select = "AIC", sle = 0.15, sls = 0.15)
```

```
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1        checks                1        1 4461.5917
2    2      employed                1        2 4372.1808
3    3       citizen                1        3  4327.698
4    4        colour                1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
               Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
checks           checks  -0.3628     0.0257 -14.1008        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 8: K- Fold Cross Validation To Find The Choose The Model With The Most Accuracy

> #We found that no matter what model selection technique we did, there are the same 4 significant independent variables for fitting our dependent variable and the model is the same model for each. Note that the lower AIC, the better!

> #But we need to know if there is still a better model where we can avoid overfitting.

> #Only looking at the AIC to determine which model you would choose is generally not enough in making sure you selected the best model.

> #We need to assess the accuracy and validity of each model to determine which model we should choose by Cross Validation or CV.

> #The technique we will use for CV is K-Fold Cross Validation because it has the advantage of using all data for estimating the model over other CV techniques.

> #Especially since over 5,000 observations and is large, we could definitely do a 10-Fold Cross Validation.

Step 9: Comparing Our Model We Got From Model Selection Techniques With A Model With the 3 Most Significant Variable

> #Since we already know that no matter what selection technique we do, we get the same model, we can try to compare the model we got with a model that has the 3 most significant independent variables.

```
forward <- glm(released  ~ checks + employed + citizen + colour, data=Arrests, family = "binomial")
```

> #looking at the forward selection output and finding the three most significant variables

```
> forward2 <- glm(released ~ checks + employed + citizen, data=Arrests, family = "binomial")
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 10: Comparing Our Models

> summary(forward)

```
Call:
glm(formula = released ~ checks + employed + citizen + colour,
    family = "binomial", data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3580   0.3579   0.4316   0.6061   1.6982

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.00474    0.12741   7.886 3.12e-15 ***
checks      -0.36283    0.02573 -14.101  < 2e-16 ***
employedYes  0.75367    0.08398   8.974  < 2e-16 ***
citizenYes   0.56839    0.09916   5.732 9.91e-09 ***
colourWhite  0.38915    0.08521   4.567 4.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4299.3  on 5221  degrees of freedom
AIC: 4309.3

Number of Fisher Scoring iterations: 5
```

> summary(forward2)

```
Call:
glm(formula = released ~ checks + employed + citizen, family = "binomial",
    data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3330   0.3689   0.4420   0.6279   1.6450

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.20519    0.11972  10.067  < 2e-16 ***
checks      -0.37653    0.02546 -14.788  < 2e-16 ***
employedYes  0.77488    0.08366   9.262  < 2e-16 ***
citizenYes   0.67335    0.09609   7.007 2.43e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4319.7  on 5222  degrees of freedom
AIC: 4327.7

Number of Fisher Scoring iterations: 5
```

># Even though the model "forward" had the lower AIC, just looking at AIC is not enough because this does not tell you if this model generally works if you applied to data outside of your training data. This is why we still need to do a K-fold Cross Validation.

Step 11: Start The CV To Choose The Model Between "Forward" and "Forward2"

```
> require(caret) #package used for cross validation
> library(e1071)
> # Define training control
> set.seed(13245)
> train.control <- trainControl(method = "cv", number = 10)
> # Train the model forward (4 predictors)
> model_forward <- train(released ~checks + employed + citizen + colour,data = Arrests, method = "glm",
+                        trControl = train.control)
> # Define training control
> set.seed(14235)
> train.control <- trainControl(method = "cv", number = 10)
> # Train the model forward2 (3 predictors)
> model_forward2 <- train(released ~ checks + employed + citizen,data = Arrests, method = "glm",
+                         trControl = train.control)
```

Source: (J. Mardekian, personal communication, April 9 , 2020)

Step 12: Cross Validating

> \# Summarize the results
> print(model_forward)

```
Generalized Linear Model

5226 samples
   4 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4704, 4703, 4704, 4704, 4704, 4703, ...
Resampling results:

  Accuracy   Kappa
  0.8279767  0.07195669
```

> \# Summarize the results
> print(model_forward2)

```
Generalized Linear Model

5226 samples
   3 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4704, 4704, 4702, 4704, 4704, 4704, ...
Resampling results:

  Accuracy   Kappa
  0.8300782  0.06043496
```

Because the accuracies are very similar between the models, we decided to select the forward model because it's kappa was higher. A kappa closer to 1 means that there is better reliability.

Source: (J. Mardekian, personal communication, April 9 , 2020)

4.

Results

Model Selection & K-Fold Cross-Validation

Full Model Selection

Included parameters	color, employed, citizen, and checks
Excluded Parameters	year, age, sex
Dependent Variable	Released (yes or no)
AIC	4315.1
P-Wald's Test/P(LR-Test)	5.56e-06, < 2e-16, 3.20e-08, < 2e-16 (<0.001)
Residual Deviance	4299.1 on 5218 d.f.
Null Deviance	4776.3 on 5225 d.f.
Log Likelihood	-2149.5327

Full Model Output Summary

```
Call:
glm(formula = released ~ colour + year + age + sex + employed +
    citizen + checks, family = "binomial", data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3909   0.3579   0.4320   0.6047   1.7067

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.371821  56.717803   0.165    0.869
colourWhite  0.389109   0.085663   4.542 5.56e-06 ***
year        -0.004218   0.028379  -0.149    0.882
age          0.002236   0.004631   0.483    0.629
sexMale      0.007317   0.150189   0.049    0.961
employedYes  0.757302   0.084735   8.937  < 2e-16 ***
citizenYes   0.576519   0.104246   5.530 3.20e-08 ***
checks      -0.364101   0.025984 -14.013  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4299.1  on 5218  degrees of freedom
AIC: 4315.1

Number of Fisher Scoring iterations: 5
```

Full Model Output Summary II

Logistic regression predicting released : Yes vs No

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
colour: White vs Black	2.11 (1.81,2.46)	1.48 (1.25,1.75)	< 0.001	< 0.001
year (cont. var.)	1.0628 (1.0093,1.1192)	0.9958 (0.9419,1.0527)	0.882	0.882
age (cont. var.)	0.987 (0.9789,0.9952)	1.0022 (0.9932,1.0114)	0.629	0.628
sex: Male vs Female	0.7907 (0.5995,1.0431)	1.0073 (0.7505,1.3521)	0.961	0.961
employed: Yes vs No	2.99 (2.56,3.5)	2.13 (1.81,2.52)	< 0.001	< 0.001
citizen: Yes vs No	2.11 (1.76,2.52)	1.78 (1.45,2.18)	< 0.001	< 0.001
checks (cont. var.)	0.65 (0.62,0.69)	0.69 (0.66,0.73)	< 0.001	< 0.001

Log-likelihood = -2149.5327
No. of observations = 5226
AIC value = 4315.0654

MODEL SELECTION CONTINUED

Forward

- Included parameters: checks, employed, citizen, color
- AIC:

Checks (1)	4461. 5
Employed (2)	4372. 1808
Citizen (3)	4327. 698.5917
Color (4)	4309. 3187

Backward

- Excluded parameters: sex, year, age
- AIC:

Sex (6)	4313. 0678
Year (5)	4311. 0896
Age (4)	4309. 3187

Bi-directional

- Included parameters: checks, employed, citizen, color
- AIC:

Checks (1)	4461. 5917
Employed (2)	4372. 1808
Citizen (3)	4327. 698
Color (4)	4309. 3187

AIC ("Akaike's Information Criterion") is a common model selection criteria which is used to measure model performance. AIC is calculated by obtaining the maximum value of the likelihood function for the model. AIC suggests adding more parameters to a model will improve the goodness of fit, but will also increase the penalty imposed by adding more predictors. In logistic regression, AIC is especially useful since it is calculated using the model's maximum likelihood estimator as a measure of fit.

Output (Forward Selection)

```
> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "backward",
+               select = "AIC", sle = 0.15, sls = 0.15)
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1                         sex  1        6 4313.0678
2    2                        year  1        5 4311.0896
3    3                         age  1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
              Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
checks           checks  -0.3628     0.0257 -14.1008        0
```

Output (Backward Selection)

```
> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "forward",
+              select = "AIC", sle = 0.15, sls = 0.15)
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1        checks                1        1 4461.5917
2    2      employed                1        2 4372.1808
3    3       citizen                1        3  4327.698
4    4        colour                1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
              Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
checks           checks  -0.3628     0.0257 -14.1008        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
```

Output (Bidirectional Selection)

```
> stepwiselogit(data=Arrests,y, exclude = NULL, include = NULL, selection = "bidirection",
+               select = "AIC", sle = 0.15, sls = 0.15)
$SummaryOfSelection
  Step EnteredEffect RemovedEffect DF NumberIn       AIC
1    1        checks                1        1 4461.5917
2    2      employed                1        2 4372.1808
3    3       citizen                1        3  4327.698
4    4        colour                1        4 4309.3187

$AnalysisOfMaximumLikelihoodEstimate
              Parameter Estimate Std. Error  z value Pr(>|z|)
(Intercept) (Intercept)   1.0047     0.1274    7.886        0
checks           checks  -0.3628     0.0257 -14.1008        0
employedYes employedYes   0.7537      0.084   8.9742        0
citizenYes   citizenYes   0.5684     0.0992   5.7323        0
colourWhite colourWhite   0.3891     0.0852   4.5668        0
```

K-Fold Cross-Validation

Our data was broken into 10 buckets, with approximately 500 observations in each bucket

4 Predictors

- Included parameters: checks, employed, citizen, color
- AIC: 4309.3
- Accuracy: 0.8279767
- Kappa: 0.07195669
- McNemar's Test:

3 Predictors

- Included parameters: checks, employed, citizen
- AIC: 4327.7
- Accuracy: 0.8300782
- Kappa: 0.06043496
- McNemar's Test:

2 Predictors

- Included parameters: checks, employed
- AIC: 4372.2
- Accuracy: 0.82797797
- Kappa: 0.02688968
- McNemar's Test:

Cross Validation (4 Predictors)

```
> summary(forward)

Call:
glm(formula = released ~ checks + employed + citizen + colour,
    family = "binomial", data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3580   0.3579   0.4316   0.6061   1.6982

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.00474    0.12741   7.886 3.12e-15 ***
checks      -0.36283    0.02573 -14.101  < 2e-16 ***
employedYes  0.75367    0.08398   8.974  < 2e-16 ***
citizenYes   0.56839    0.09916   5.732 9.91e-09 ***
colourWhite  0.38915    0.08521   4.567 4.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4299.3  on 5221  degrees of freedom
AIC: 4309.3

Number of Fisher Scoring iterations: 5
```

```
> print(model_forward)
Generalized Linear Model

5226 samples
   4 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4704, 4703, 4704, 4704, 4704, 4703, ...
Resampling results:

  Accuracy   Kappa
  0.8279767  0.07195669
```

Cross Validation (3 Predictors)

```
> summary(forward2)

Call:
glm(formula = released ~ checks + employed + citizen, family = "binomial",
    data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3330   0.3689   0.4420   0.6279   1.6450

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.20519    0.11972  10.067  < 2e-16 ***
checks      -0.37653    0.02546 -14.788  < 2e-16 ***
employedYes  0.77488    0.08366   9.262  < 2e-16 ***
citizenYes   0.67335    0.09609   7.007 2.43e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4319.7  on 5222  degrees of freedom
AIC: 4327.7

Number of Fisher Scoring iterations: 5
```

```
> print(model_forward2)
Generalized Linear Model

5226 samples
   3 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4704, 4704, 4702, 4704, 4704, 4704, ...
Resampling results:

  Accuracy   Kappa
  0.8300782  0.06043496
```

Cross Validation (2 Predictors)

```
> summary(forward3)

Call:
glm(formula = released ~ checks + employed, family = "binomial",
    data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2886   0.3890   0.4657   0.6599   1.4072

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.73366    0.09510  18.231   <2e-16 ***
checks      -0.37655    0.02536 -14.850   <2e-16 ***
employedYes  0.80953    0.08305   9.748   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4366.2  on 5223  degrees of freedom
AIC: 4372.2

Number of Fisher Scoring iterations: 5
```

```
> print(model_forward3)
Generalized Linear Model

5226 samples
   2 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4704, 4704, 4702, 4704, 4704, 4704, ...
Resampling results:

  Accuracy   Kappa
  0.8279797  0.02688968
```

5.

Discussion

Chosen Model & Explanation of Chosen Model

Final Model Overview

CHOSEN MODEL: Forward Model

EXPLANATION: Model selection yielded the same result between forward, backward, and bi-directional. However, cross-validation analysis, showed forward modeling to be most effective due to higher kappa value. Thus, we chose forward modeling as the best model for our data.

LIMITATIONS: Due to limited time of this project, we acknowledge and understand that some of our methods could have been different if feasibility and timing were not at stake. In real world application, data should have been split, and the training and test datasets would both have k-fold cross-validation performed upon them.

6.

Works Cited

Drug Policy Alliance. (n.d.). *Drug War Statistics.* Retrieved April 10, 2020 from https://www.drugpolicy.org/issues/drug-war-statistics

Friendly, M. (n.d.). Arrests for Marijuana Possession. [Dataset]. York University. http://math.furman.edu/~dcs/courses/math47/R/library/effects/html/Arrests.html

Gramlich, J. (2020). *Four-in-ten U.S. drug arrests in 2018 were for marijuana offenses - mostly possession*. Pew Research Center. Retrieved March 31, 2020, from https://www.pewresearch.org/fact-tank/2020/01/22/four-in-ten-u-s-drug-arrests-in-2018-were-for-marijuana-offenses-mostly-possession/

Hoover, A. (2019, Dec 6). *Legal weed is now up to N.J. voters as lawmakers vote to put it on 2020 ballot*. NJ.com. Retrieved April 10, 2020 from https://www.nj.com/marijuana/2019/12/voters-could-decide-if-nj-will-legalize-weed-after-senate-votes-to-put-it-on-2020-ballot.html

Kelly, M. (2019, June 26). *How many people are in prison on marijuana charges?*. The Washington Post. Retrieved April, 10, 2020, from https://www.washingtonpost.com/politics/2019/live-updates/general-election/fact-checking-the-first-democratic-debate/how-many-people-are-in-prison-on-marijuana-charges/?arc404=true

The State of New Jersey. (2019, Nov 26). *Statement by Governor Murphy on Marijuana Decriminalization*. Official Site of the State of New Jersey. Retrieved April 10, 2020 from https://www.nj.gov/governor/news/news/562019/approved/20191126b.shtml

Wagner, P., Sakala, L., & Begley, J. (n.d.). *States of Incarceration: The Global Context*. Prison Policy Initiative. Retrieved April 10, 2020 from https://www.prisonpolicy.org/global/

(J. Mardekian, personal communication, April 9 , 2020)