MATH4939 Project Arrest Dataset

Team 3

09/04/2020

Introduction

"Arrests" dataset is based on the police treatment of individuals arrested in Toronto for possession of small amounts of marijuana from 1997 to 2002. The dataset is just a part of a large data set mentioned in a series of articles in the Toronto star. The dataset contains 5226 observations with 8 variables as below.

released: whether or not the person who is arrested is released with a summon (Yes or No) xcolour: The arrested persons race (Black or White) year: 1997 - 2002 age: The age of the arrested person in years sex: Gender of the arrested person (Male or Female) employed: Is the arrested person employed (Yes or No) citizen: Is the person a citizen of toronto (Yes or No) checks: Number obtained from the police databases (of previous arrests, previous conviction, parole status, etc.) the arrested persons name appeared upon labeled from 1 to 6

According to the dataset, the variable "released" is the independent variable y, and the rest are dependent variables. In this project, we will build two models using logistic regression method, compare the two models in different ways, and find out the factors which can influence the independent variable "released" significantly in order to explore the patterns of discrimination in the dataset.

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

Motivation

We will build two models for the dataset. The first model is a simple model, which is :

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

. Another model is more complicated, which contains the interaction terms and quadratic terms. We are going to compare these two models in different ways and use the better one to find out the potential patterns of discrimination.

Model Selection

Based on the given data set we build an interaction model which will be our base.

```
##
## Call:  glm(formula = released ~ (colour + year + age + sex + citizen +
##     employed + checks)^2, family = "binomial", data = Arrests)
##
## Coefficients:
##             (Intercept)              colourWhite                     year
##              -2.279e+02                3.944e+02                1.137e-01
##                     age                  sexMale               citizenYes
##               5.637e+00               -5.714e+02               -1.453e+02
##             employedYes                   checks         colourWhite:year
##               3.164e+02                1.193e+02                -1.966e-01
##         colourWhite:age      colourWhite:sexMale   colourWhite:citizenYes
##              -3.652e-02                1.021e-01                5.424e-02
## colourWhite:employedYes       colourWhite:checks                 year:age
##               7.222e-03               -2.715e-02               -2.793e-03
##            year:sexMale          year:citizenYes         year:employedYes
##               2.860e-01                7.268e-02               -1.572e-01
##             year:checks              age:sexMale           age:citizenYes
##              -5.976e-02               -2.149e-02                8.812e-03
##         age:employedYes               age:checks       sexMale:citizenYes
##              -2.170e-02                1.565e-03                6.736e-01
##     sexMale:employedYes           sexMale:checks   citizenYes:employedYes
##              -8.135e-01               -1.232e-01                7.732e-02
##       citizenYes:checks       employedYes:checks
##              -1.479e-01               -3.319e-02
##
## Degrees of Freedom: 5225 Total (i.e. Null);  5197 Residual
## Null Deviance:       4776
## Residual Deviance: 4229  AIC: 4287
```

After evaluating the model using wald test we see that there are some p-values in the predictors that may cause problems in further analysis.

```
##  numDF denDF  F.value p.value
##     29  5197 58.14915 <.00001
##                          Estimate    Std.Error  DF   t-value   p-value
## (Intercept)             -227.911234 327.888242 5197 -0.695088 0.48703
## colourWhite              394.375896 123.758087 5197  3.186668 0.00145
## year                       0.113709   0.164072 5197  0.693043 0.48831
## age                        5.636549   6.828481 5197  0.825447 0.40916
## sexMale                 -571.446573 247.518401 5197 -2.308703 0.02100
## citizenYes              -145.316319 131.011546 5197 -1.109187 0.26740
## employedYes              316.376704 123.560675 5197  2.560497 0.01048
## checks                   119.348085  38.799133 5197  3.076050 0.00211
## colourWhite:year          -0.196650   0.061930 5197 -3.175348 0.00151
## colourWhite:age           -0.036523   0.010824 5197 -3.374135 0.00075
## colourWhite:sexMale        0.102070   0.396608 5197  0.257358 0.79691
## colourWhite:citizenYes     0.054240   0.215764 5197  0.251384 0.80153
## colourWhite:employedYes    0.007222   0.184686 5197  0.039102 0.96881
```

```
## colourWhite:checks        -0.027151   0.058996 5197 -0.460218 0.64538
## year:age                  -0.002793   0.003416 5197 -0.817394 0.41374
## year:sexMale               0.286046   0.123866 5197  2.309325 0.02096
## year:citizenYes            0.072681   0.065533 5197  1.109067 0.26745
## year:employedYes          -0.157201   0.061826 5197 -2.542662 0.01103
## year:checks               -0.059756   0.019414 5197 -3.078048 0.00209
## age:sexMale               -0.021493   0.019366 5197 -1.109845 0.26712
## age:citizenYes             0.008812   0.012281 5197  0.717520 0.47309
## age:employedYes           -0.021700   0.010302 5197 -2.106314 0.03522
## age:checks                 0.001565   0.003260 5197  0.479990 0.63125
## sexMale:citizenYes         0.673569   0.588149 5197  1.145235 0.25216
## sexMale:employedYes       -0.813452   0.350712 5197 -2.319430 0.02041
## sexMale:checks            -0.123188   0.105853 5197 -1.163763 0.24457
## citizenYes:employedYes     0.077319   0.225704 5197  0.342568 0.73194
## citizenYes:checks         -0.147893   0.067658 5197 -2.185886 0.02887
## employedYes:checks        -0.033193   0.058290 5197 -0.569457 0.56907
##                          Lower 0.95   Upper 0.95
## (Intercept)              -870.710085 414.887617
## colourWhite               151.757998 636.993793
## year                       -0.207942   0.435360
## age                        -7.750145  19.023242
## sexMale                 -1056.686735 -86.206411
## citizenYes               -402.154047 111.521409
## employedYes                74.145816 558.607592
## checks                     43.285467 195.410704
## colourWhite:year           -0.318059  -0.075241
## colourWhite:age            -0.057743  -0.015303
## colourWhite:sexMale        -0.675449   0.879589
## colourWhite:citizenYes     -0.368749   0.477228
## colourWhite:employedYes    -0.354841   0.369284
## colourWhite:checks         -0.142808   0.088506
## year:age                   -0.009490   0.003905
## year:sexMale                0.043217   0.528875
## year:citizenYes            -0.055792   0.201154
## year:employedYes           -0.278405  -0.035997
## year:checks                -0.097815  -0.021697
## age:sexMale                -0.059458   0.016472
## age:citizenYes             -0.015264   0.032887
## age:employedYes            -0.041897  -0.001503
## age:checks                 -0.004826   0.007956
## sexMale:citizenYes         -0.479451   1.826588
## sexMale:employedYes        -1.500994  -0.125909
## sexMale:checks             -0.330704   0.084328
## citizenYes:employedYes     -0.365156   0.519793
## citizenYes:checks          -0.280531  -0.015255
## employedYes:checks         -0.147466   0.081079
```

Thus we run stepwise regression to obtain the selected model which we will use to compare with the additative model.

Based on the stepwise regression we end up with our selected model which is

$$logit(\pi(released)) = \beta_0+\beta_1 colour+\beta_2 year+\beta_3 age+\beta_4 sex+\beta_5 employed+\beta_6 citizen+\beta_7 checks+\beta_8 colour*year + \beta_9 colour*age+\beta_{10} year*sex+\beta_{11} year*employed+\beta_{12} year*checks+ \beta_{13} age*employed+\beta_{14} sex*employed+\beta_{15} citizen*checks$$

```
##
## Call:  glm(formula = released ~ colour + year + age + sex + citizen +
##     employed + checks + colour * year + colour * age + year *
##     sex + year * employed + year * checks + age * employed +
##     sex * employed + citizen * checks, family = "binomial", data = Arrests)
##
## Coefficients:
##         (Intercept)          colourWhite                 year
##           -165.51999             366.20168              0.08246
##                 age              sexMale           citizenYes
##             0.04449            -571.63777              0.92475
##         employedYes                checks     colourWhite:year
##           313.81186             114.39688             -0.18254
##     colourWhite:age         year:sexMale     year:employedYes
##            -0.03434               0.28606             -0.15598
##         year:checks      age:employedYes  sexMale:employedYes
##            -0.05734              -0.02598             -0.59402
##   citizenYes:checks
##            -0.15067
##
## Degrees of Freedom: 5225 Total (i.e. Null);  5210 Residual
## Null Deviance:       4776
## Residual Deviance: 4237  AIC: 4269
```

Model Comparison

We compared the simple additive model

$$logit(\pi(released)) = \beta_0 + \beta_1 colour + \beta_2 year + \beta_3 age + \beta_4 sex + \beta_5 employed + \beta_6 citizen + \beta_7 checks$$

with the model we get after selection in different ways, which is

$$logit(\pi(released)) = \beta_0 + \beta_1 colour + \beta_2 year + \beta_3 age + \beta_4 sex + \beta_5 employed + \beta_6 citizen + \beta_7 checks+ \beta_8 colour * year + \beta_9 colour * age+ \beta_{10} year*sex+\beta_{11} year*employed+\beta_{12} year*checks+\beta_{13} age*employed+\beta_{14} sex*employed+\beta_{15} citizen*checks$$

1.Comparison in Overall Model Significance

We did wald test for the two models to compare their model significance.

For the Additive Model

```
##  numDF denDF  F.value p.value
##      8  5218 217.7688 <.00001
```

```
##              Estimate  Std.Error DF   t-value    p-value Lower 0.95  Upper 0.95
## (Intercept)  9.371821 56.717803 5218   0.165236 0.86876 -101.818821 120.562463
## colourWhite  0.389109  0.085663 5218   4.542295 0.00001    0.221172   0.557045
## year        -0.004218  0.028379 5218  -0.148650 0.88184   -0.059853   0.051416
## age          0.002236  0.004631 5218   0.482709 0.62932   -0.006844   0.011315
## sexMale      0.007317  0.150189 5218   0.048716 0.96115   -0.287118   0.301751
## citizenYes   0.576519  0.104246 5218   5.530360 <.00001    0.372153   0.780886
## employedYes  0.757302  0.084735 5218   8.937291 <.00001    0.591186   0.923418
## checks      -0.364101  0.025984 5218 -14.012732 <.00001   -0.415040  -0.313162
```

For the Selected Model

```
##  numDF denDF  F.value p.value
##     16  5210 106.4882 <.00001
##                      Estimate    Std.Error  DF   t-value   p-value Lower 0.95
## (Intercept)         -165.519991 260.239290 5210 -0.636030 0.52478  -675.698148
## colourWhite          366.201678 115.884048 5210  3.160070 0.00159   139.020339
## year                   0.082462   0.130160 5210  0.633541 0.52641    -0.172706
## age                    0.044486   0.010409 5210  4.273954 0.00002     0.024081
## sexMale             -571.637770 238.882970 5210 -2.392962 0.01675 -1039.948583
## citizenYes             0.924749   0.170189 5210  5.433654 <.00001     0.591107
## employedYes          313.811861 117.768020 5210  2.664661 0.00773    82.937148
## checks               114.396884  38.215283 5210  2.993485 0.00277    39.478902
## colourWhite:year      -0.182539   0.057958 5210 -3.149516 0.00164    -0.296160
## colourWhite:age       -0.034340   0.010128 5210 -3.390676 0.00070    -0.054195
## year:sexMale           0.286062   0.119464 5210  2.394540 0.01668     0.051862
## year:employedYes      -0.155985   0.058895 5210 -2.648520 0.00811    -0.271443
## year:checks           -0.057339   0.019121 5210 -2.998709 0.00272    -0.094824
## age:employedYes       -0.025980   0.009828 5210 -2.643335 0.00823    -0.045247
## sexMale:employedYes   -0.594024   0.312610 5210 -1.900204 0.05746    -1.206871
## citizenYes:checks     -0.150671   0.063422 5210 -2.375702 0.01755    -0.275004
##                      Upper 0.95
## (Intercept)          344.658166
## colourWhite          593.383017
## year                   0.337629
## age                    0.064892
## sexMale             -103.326957
## citizenYes             1.258392
## employedYes          544.686574
## checks               189.314866
## colourWhite:year      -0.068917
## colourWhite:age       -0.014485
## year:sexMale           0.520263
## year:employedYes      -0.040526
## year:checks           -0.019853
## age:employedYes       -0.006712
## sexMale:employedYes    0.018824
## citizenYes:checks     -0.026338
```

Comparing these two outputs, we found that the overall p-values in both models are small, which means these two models are both significant. For the model we selected, the p-values for the variable year, age and sex are smaller that the p-values of additive model, which means these factors become more significant in the model we selected. What's more, in the model we selected, p-values are very small for most of the interaction terms. Thus, we can conclude that the significance gets improved using the model we selected.

2. Comparision in ANOVA

```
## Analysis of Deviance Table
##
## Model 1: released ~ colour + year + age + sex + citizen + employed + checks
## Model 2: released ~ colour + year + age + sex + citizen + employed + checks +
##     colour * year + colour * age + year * sex + year * employed +
##     year * checks + age * employed + sex * employed + citizen *
##     checks
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5218     4299.1
## 2      5210     4236.5  8   62.525 1.487e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the ANOVA table, the residual deviance of model 2 is smaller than that of model 1, which means the model we selected is better than the additive model.

3.Comparison in AIC

```
##        df      AIC
## model2 16 4268.541
## fit     8 4315.065
```

According to the output of AIC, the value of AIC in the model we selected is much smaller than the additive model. Thus, in this aspect, the model we selected is better than the additive model.

Model Evaluation

1. Comparison in Multicollinearity (VIF)

```
##         colour           year            age            sex        citizen
##   1.959944e+06   2.280712e+01   5.165030e+00   2.397521e+06   3.159094e+00
##       employed         checks    colour:year     colour:age       year:sex
##   2.015186e+06   2.270709e+06   1.960002e+06   1.259478e+01   2.397105e+06
##  year:employed    year:checks   age:employed   sex:employed citizen:checks
##   2.014948e+06   2.272626e+06   1.129668e+01   1.483983e+01   7.154602e+00
```

In the selected model, the VIF score for the predictor variables have some predictors that are higher than 10, we can say that there is some form of multicollinearity that exists in this model. We see that since our model contains alot of interaction terms we concluded that VIF scores are high because of the amount of interaction terms we have in the model and waved off the multicollinearity theory for our selected model.

2. Prediction Performance Evaluation

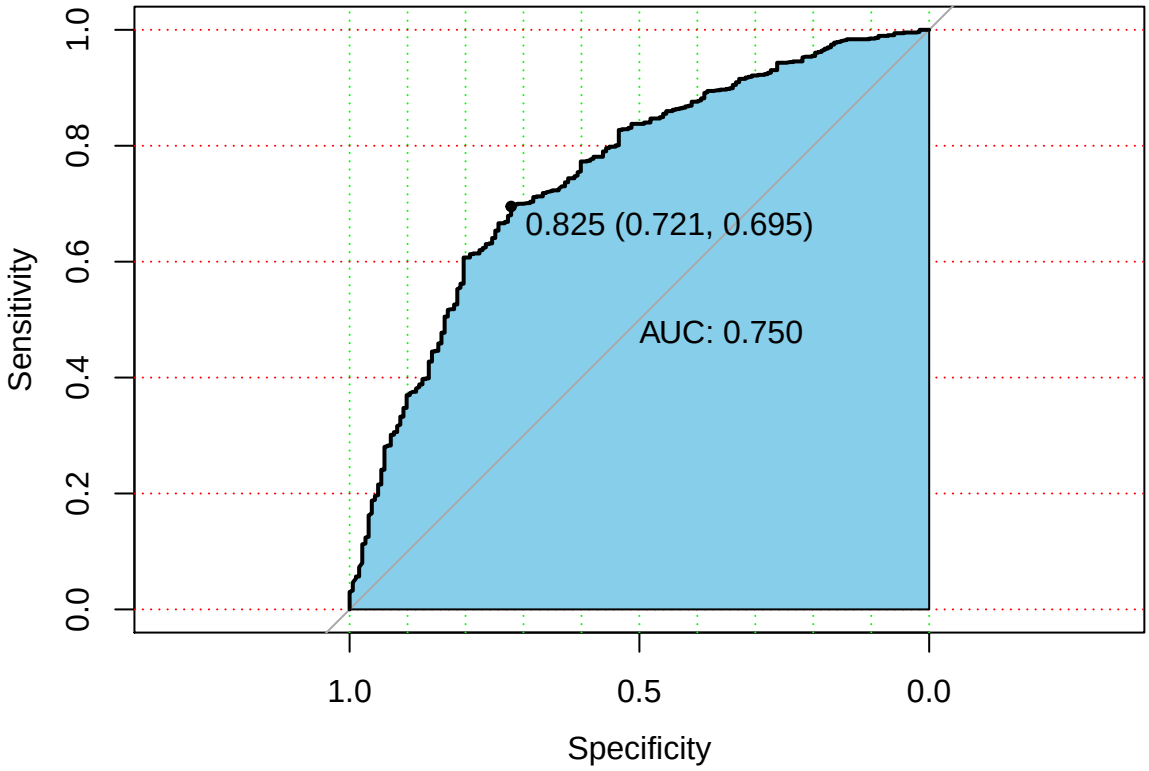
```
## [1] 0.8339713
```

```
## [1] 0.8307839
```

```
## [1] 0.8327593
```

```
## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```



Based on the best logistic model selected, using a random split 80% as training data and left 20% as testing data, the logistic model has an about 83.4% accuracy in predicting released status in training set and about 83.1% accuracy in predicting released status in testing set, and about 83.3% accuracy in predicting released status using whole data. This shows the best logistic model is consistent in prediction performance, the inferences based on the best logistic model is reliable which means given new data, the model would also performs well compare with a random guess with 50% accuracy only. The AUC on testing set is 0.75 which is high, so the model prediction performance is well

Model Analysis

Interpretation of betai in Selected Prediction Model

```
##         (Intercept)         colourWhite                year                 age
##       -165.51999080        366.20167820          0.08246162          0.04448623
##             sexMale          citizenYes         employedYes              checks
##       -571.63776988          0.92474947        313.81186106        114.39688397
##    colourWhite:year     colourWhite:age         year:sexMale    year:employedYes
##         -0.18253853         -0.03433990          0.28606233         -0.15598455
##         year:checks     age:employedYes sexMale:employedYes   citizenYes:checks
##         -0.05733875         -0.02597960         -0.59402355         -0.15067117
```

Since the model we selected is :

$$logit(\pi(released)) = \beta_0 + \beta_1 colour + \beta_2 year + \beta_3 age + \beta_4 sex + \beta_5 employed + \beta_6 citizen + \beta_7 checks +$$
$$\beta_8 colour * year + \beta_9 colour * age +$$
$$\beta_{10} year{*}sex + \beta_{11} year{*}employed + \beta_{12} year{*}checks + \beta_{13} age{*}employed + \beta_{14} sex{*}employed + \beta_{15} citizen{*}checks$$

According to the output, β_1 = 366.201678, β_2= 0.082462, β_3 = 0.044486, β_4= -571.63776988 , β_5= 0.92474947, β_6= 313.81186106 , β_7= 114.39688397, β_8= -0.18253853, β_9= -0.03433990, β_{10}= 0.28606233, β_{11}= -0.15598455, β_{12}= -0.05733875, β_{13}= -0.02597960, β_{14}= -0.59402355, β_{15}= -0.15067117. To interpret β_i, we can take partial derivative of $logit(\pi(released))$ with respect to independent variables.

For example, if we take partial derivative with respect to 'colour', we will get :

$$\frac{dlogit(\pi(released))}{dcolour} = \beta_1 + \beta_8 * year + \beta_9 * age$$

Thus beta1 is the partial derivative of log odds of dependent variable 'released' with respect to 'colour' when year=age=0.

Similarly, if we take partial derivative with respect to 'year', we will get :

$$\frac{dlogit(\pi(released))}{dyear} = \beta_2 + \beta_8 * colour + \beta_{10} * sex + \beta_{11} * employed + \beta_{12} * checks$$

Thus β_2 is the partial derivative of log odds of dependent variable 'released' with respect to 'year' when colour=sex=employed=checks=0. For other β_i , we also take partial derivative with respect to each independent variable, since the interpretation methods are similar, we do not enter into details here.

Wald Test for Some Predictor Variables in Selected Model

1. Wald Test of Terms Involving 'colour'

```
##        numDF denDF F.value p.value
## colour     3  5210 14.1791 <.00001
##                   Estimate   Std.Error  DF   t-value   p-value Lower 0.95
## colourWhite      366.201678 115.884048 5210  3.160070 0.00159 139.020339
## colourWhite:year  -0.182539   0.057958 5210 -3.149516 0.00164  -0.296160
## colourWhite:age   -0.034340   0.010128 5210 -3.390676 0.00070  -0.054195
##                   Upper 0.95
## colourWhite      593.383017
## colourWhite:year  -0.068917
## colourWhite:age   -0.014485
```

We did wald test for all the terms involving the 'colour' variable. We can see from the output that p-values for all terms which involve 'colour' are small, which means they are all significant. Thus we can conclude that the variable 'colour' influences the response variable 'released' significantly. It seems that there exists racial discrimination as for this data set.

2. Wald Test of Terms Involving 'year'

```
##      numDF denDF  F.value p.value
## year     5  5210 5.292712   7e-05
##                   Estimate  Std.Error DF   t-value   p-value Lower 0.95
```

```
## year              0.082462 0.130160  5210  0.633541 0.52641 -0.172706
## colourWhite:year -0.182539 0.057958  5210 -3.149516 0.00164 -0.296160
## year:sexMale      0.286062 0.119464  5210  2.394540 0.01668  0.051862
## year:employedYes -0.155985 0.058895  5210 -2.648520 0.00811 -0.271443
## year:checks      -0.057339 0.019121  5210 -2.998709 0.00272 -0.094824
##                  Upper 0.95
## year              0.337629
## colourWhite:year -0.068917
## year:sexMale      0.520263
## year:employedYes -0.040526
## year:checks      -0.019853
```

We are interested in the variable 'year' since the p-value is very large in the additive model which is equal to 0.88184. But in the selected model, p-value is smaller, which is equal to 0.52641. Thus the significance of variable 'year' is enhanced after we add interaction terms between 'year' and other variables. We can see from the output that all the interaction terms including 'year' are significant with small p-values. In this way, we can conclude that although 'year' individually does not have a strong relationship with 'released', the terms where 'year' interacts with other variables are important when we do the analysis. Therefore, we can analyze the relationship between predictor variables such as colour and the response variable 'released' in different years.

3.Wald Test of Terms Involving 'age'

```
##     numDF denDF  F.value p.value
## age     3  5210 6.799709 0.00014
##                 Estimate  Std.Error DF   t-value   p-value Lower 0.95
## age              0.044486 0.010409  5210  4.273954 0.00002  0.024081
## colourWhite:age -0.034340 0.010128  5210 -3.390676 0.00070 -0.054195
## age:employedYes -0.025980 0.009828  5210 -2.643335 0.00823 -0.045247
##                 Upper 0.95
## age              0.064892
## colourWhite:age -0.014485
## age:employedYes -0.006712
```

We did wald test for all the terms involving the age variable. We can see from the output that p-values for all terms which involve age are small, which means they are all significant. The interesting thing is the factor 'age' is not significant in additive model with p-value is equal to 0.62932. Thus probably the interaction between age and other variables enhances the significance of age. So we can focus on the interaction of age and other predictor factors

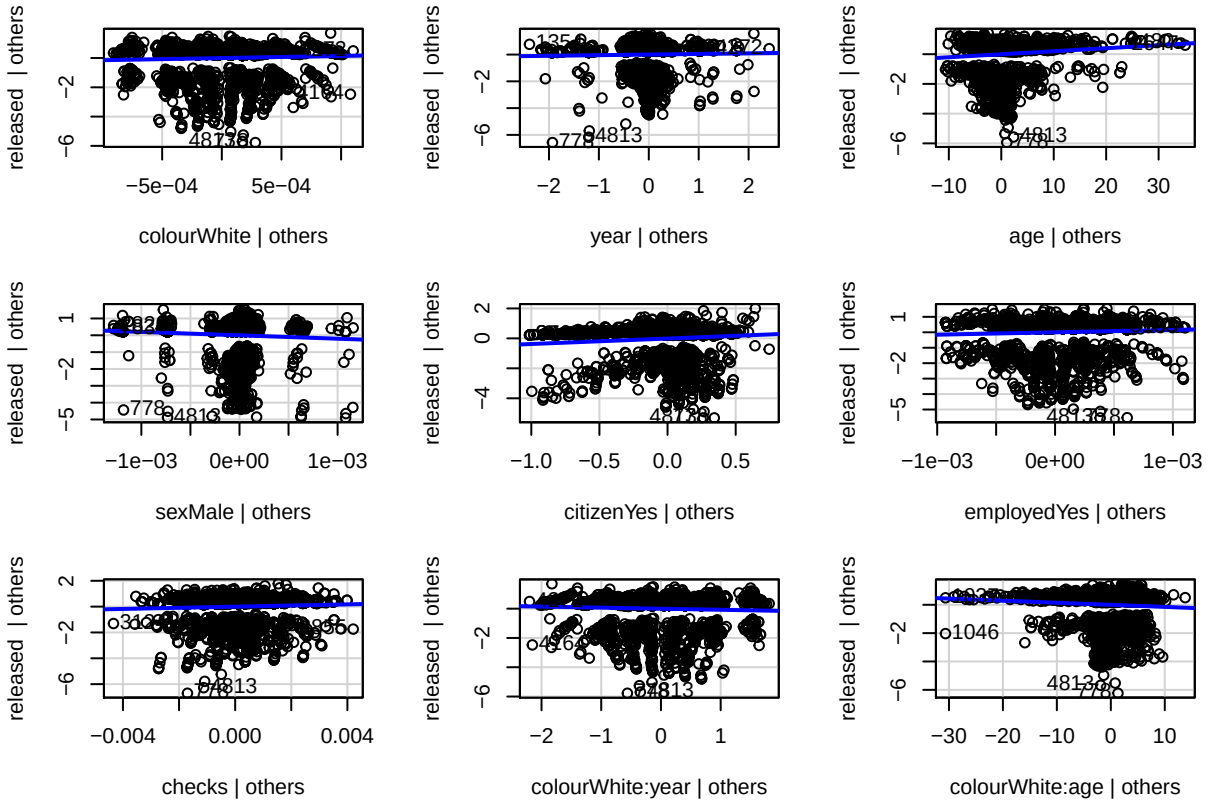
4. Wald Test of All Interaction Terms

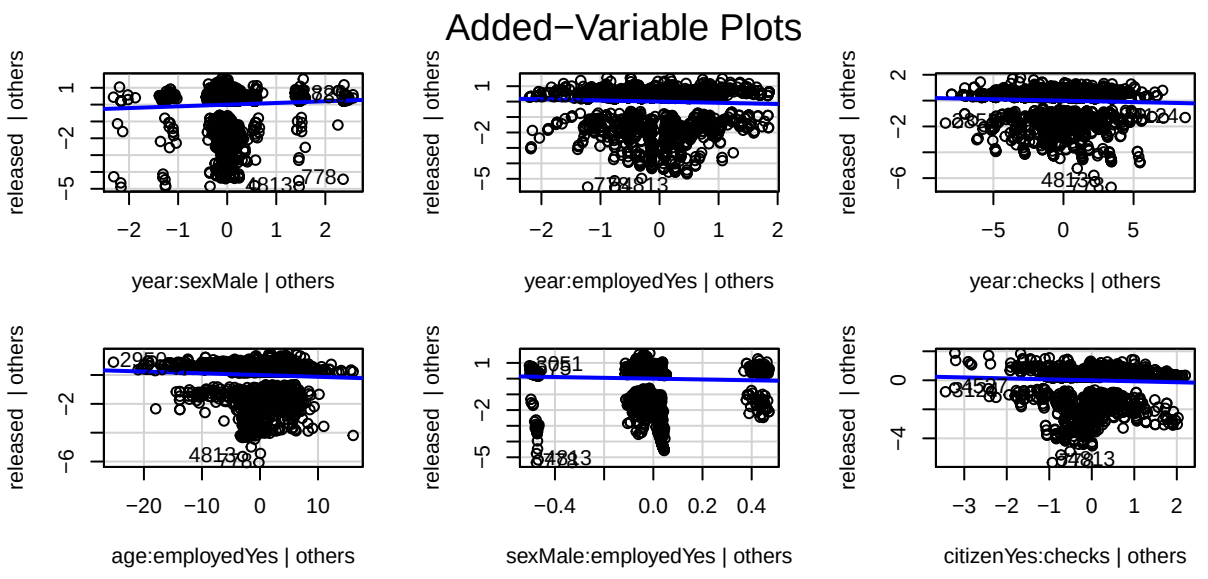
```
##   numDF denDF  F.value p.value
## :     8  5210 7.543122 <.00001
##                     Estimate  Std.Error DF   t-value   p-value Lower 0.95
## colourWhite:year    -0.182539 0.057958  5210 -3.149516 0.00164 -0.296160
## colourWhite:age     -0.034340 0.010128  5210 -3.390676 0.00070 -0.054195
## year:sexMale         0.286062 0.119464  5210  2.394540 0.01668  0.051862
## year:employedYes    -0.155985 0.058895  5210 -2.648520 0.00811 -0.271443
## year:checks         -0.057339 0.019121  5210 -2.998709 0.00272 -0.094824
## age:employedYes     -0.025980 0.009828  5210 -2.643335 0.00823 -0.045247
## sexMale:employedYes -0.594024 0.312610  5210 -1.900204 0.05746 -1.206871
```

```
## citizenYes:checks   -0.150671 0.063422  5210 -2.375702 0.01755 -0.275004
##                     Upper 0.95
## colourWhite:year    -0.068917
## colourWhite:age     -0.014485
## year:sexMale         0.520263
## year:employedYes    -0.040526
## year:checks         -0.019853
## age:employedYes     -0.006712
## sexMale:employedYes  0.018824
## citizenYes:checks   -0.026338
```

According to the output of wald test for all interaction terms, although for some terms the p-values are not that small, the overall p-value for interaction terms are really small. Thus to involve interaction terms in the model is still meaningful, we can choose some of the significant interaction terms to explore their relationships with response variable.

Added Variable plot





The Added Variable plot helps us evaluate the residuals and coefficients of the predictors while holding other variables constant. In other words it helps us answer the question if the effect of a particular predictor on the dependant variable while holding other predictors constant

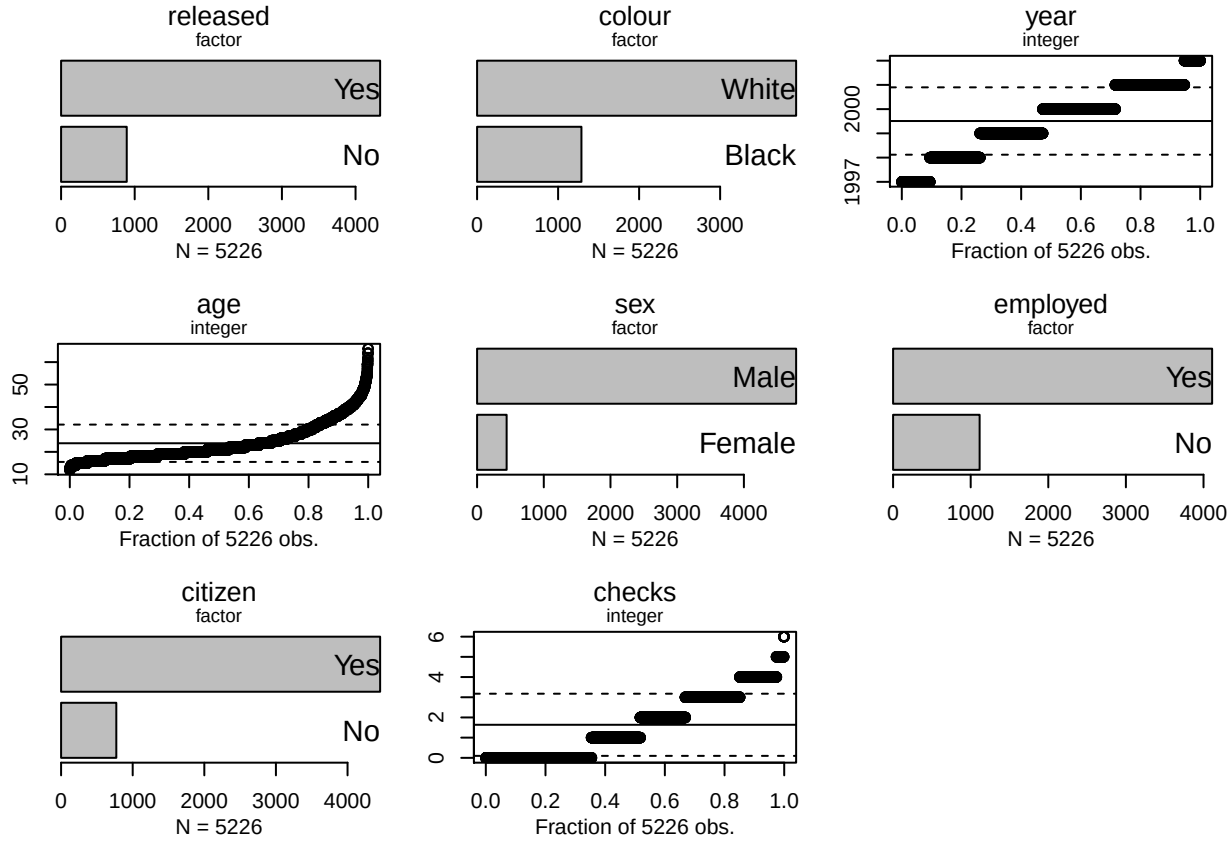
These plots show the value of each predictor after other predictors are accounted for on the X-axis and value of the dependent variable after other predictors are accounted for on the Y-axis

We can see which predictor has the strongest influence on dependant variable (released), after accounting for all other predictors, this is done by evaluating the slope of each chart in the grid

After analyzing and calculating the slopes of each grid. I come to a conclusion claiming that the interaction term based on colour have the strongest influence out of all the predictor, especially colour*year.

Data Visualization and Analysis

Xq plot



From The Xq plots, we can get basic information about the data set. For example, there are much more males than females. In addition, the number of white people is more than twice the number of black people. The number of people who are employed is also far more than those who are not employed

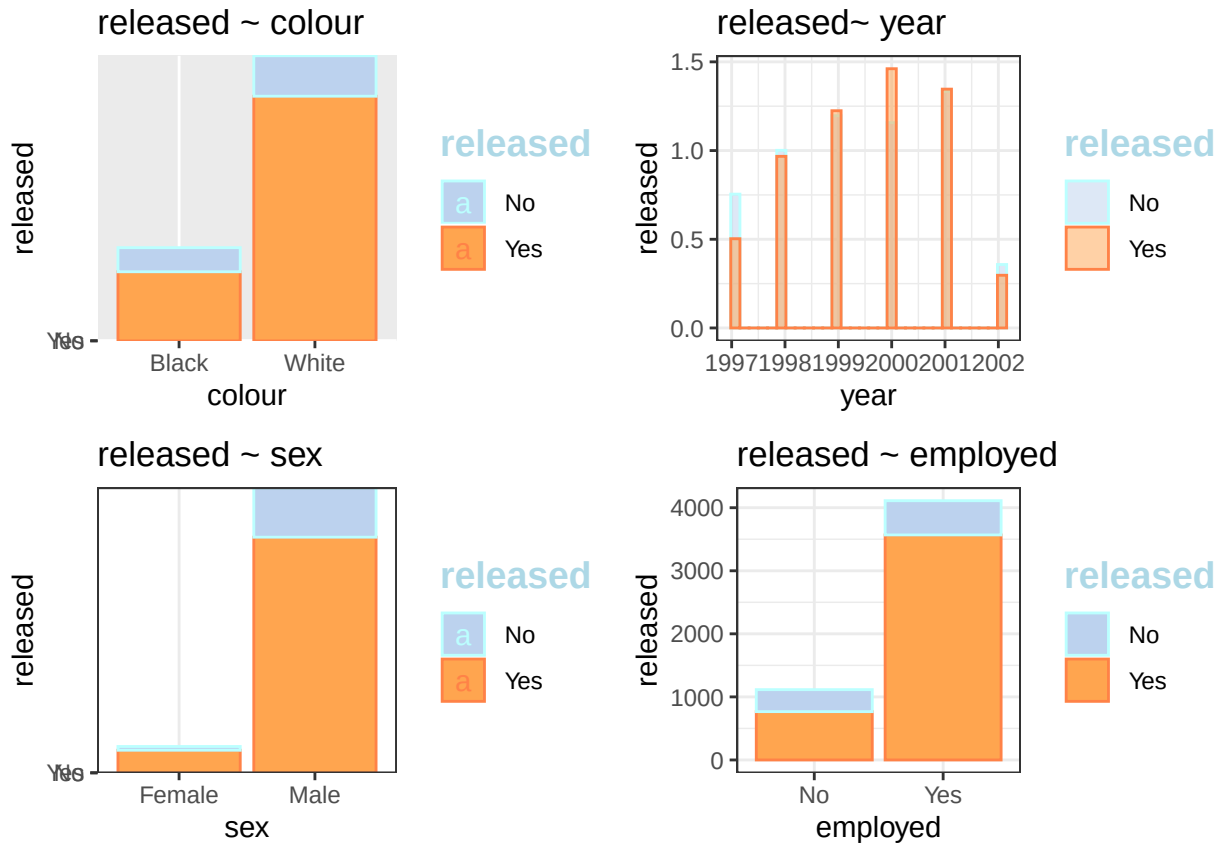
Visualization plot

In the color and released histogram, it is clear that color is a very important factor. The reason is that whites are released in significantly higher Numbers than blacks. We have to suspect that there is racial discrimination. Similarly, whites arrest fewer people than blacks. And more whites own marijuana than blacks do.

As can be seen from the figure, the number of arrests and releases increased from 1997 to 2000, reaching a peak in 2000 and the second highest in 2001. My wild guess is that it may have something to do with the 2000 U.S. presidential election and the events of 9/11. Some Canadians were unhappy with the Bush presidency and began to possess marijuana, as well as the damage to the global economy and fears about security caused by 9/11. Second, it is clear that most of those arrested each year are released. In 2002, marijuana possession was brought under control.

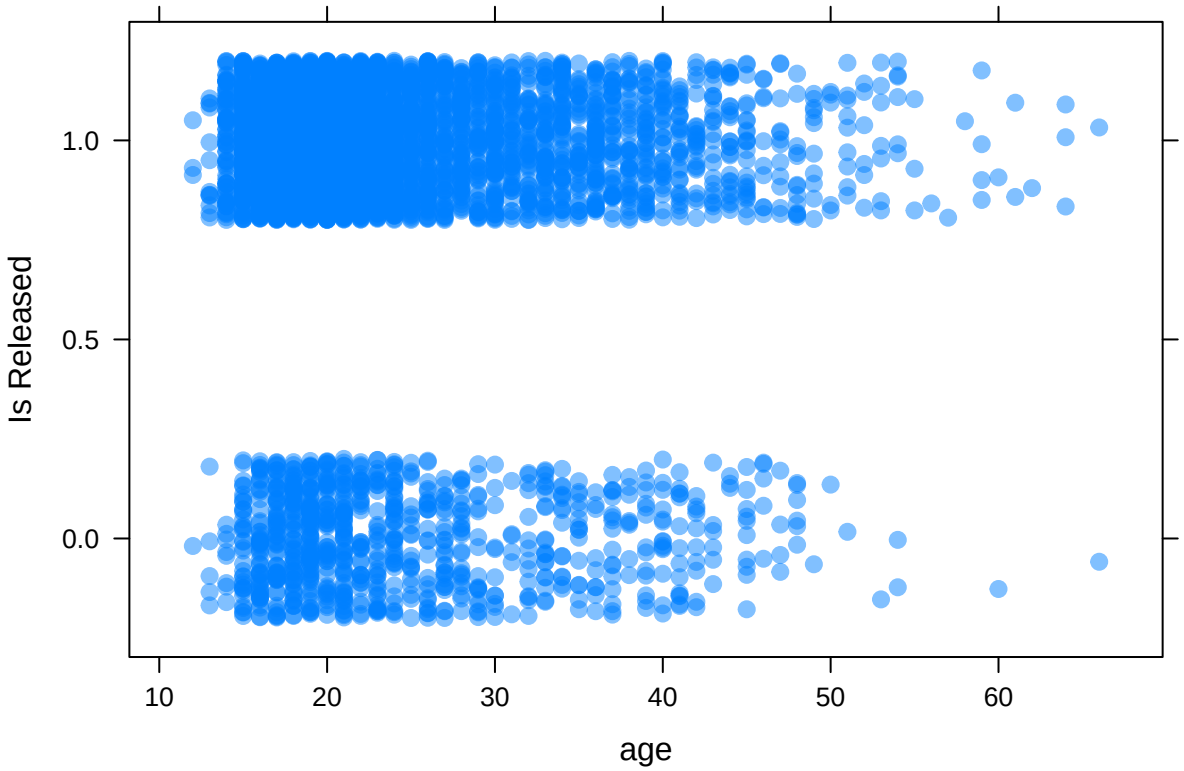
From the figure, it is obvious that the arrest rate of men is much higher than that of women. This may have to do with the fact that men are more stressed than women. In general, men's jobs are more challenging than women's, so they may need marijuana to relax.

As can be seen from the histogram, people with jobs are more likely to be released. Because people who are employed are more likely to give up marijuana than people who aren't.

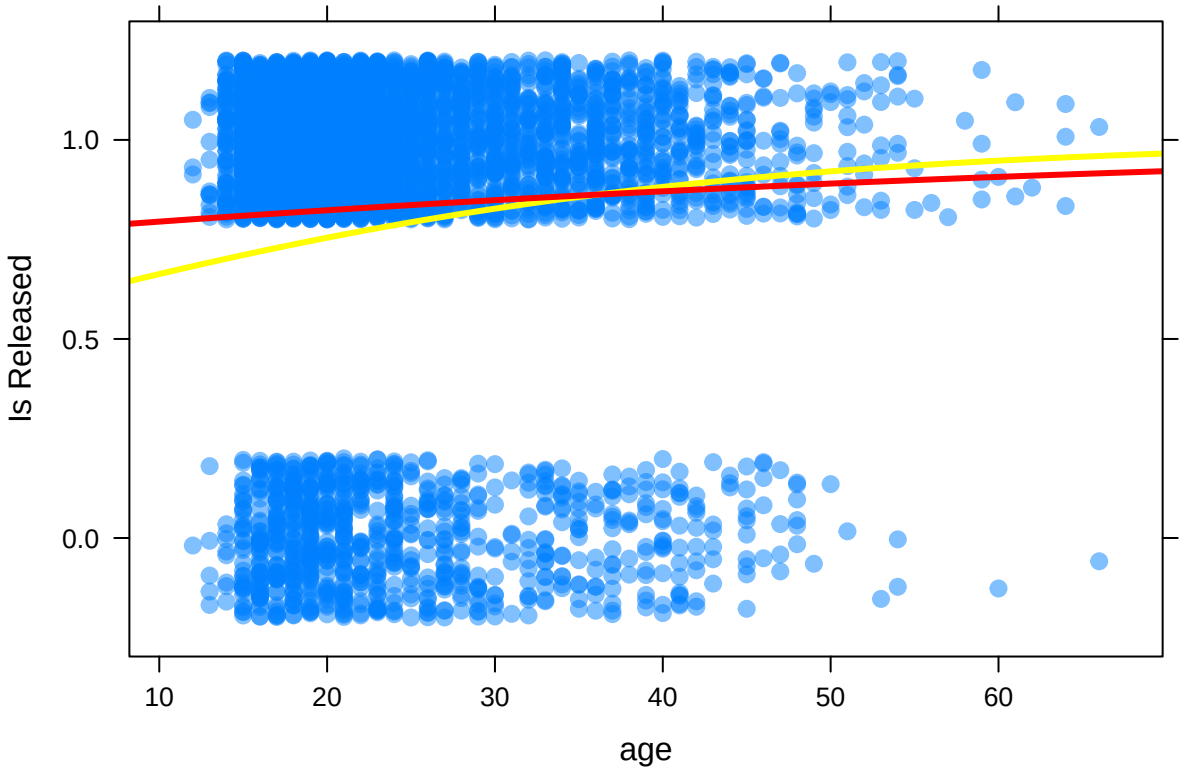


With ggplot, we can only make a basic observation and prediction of the data. More detailed analysis, we also need to build a model, more complex analysis. And each variable may affect the other. Could, for example, employed white Toronto residents be the most likely to be released?

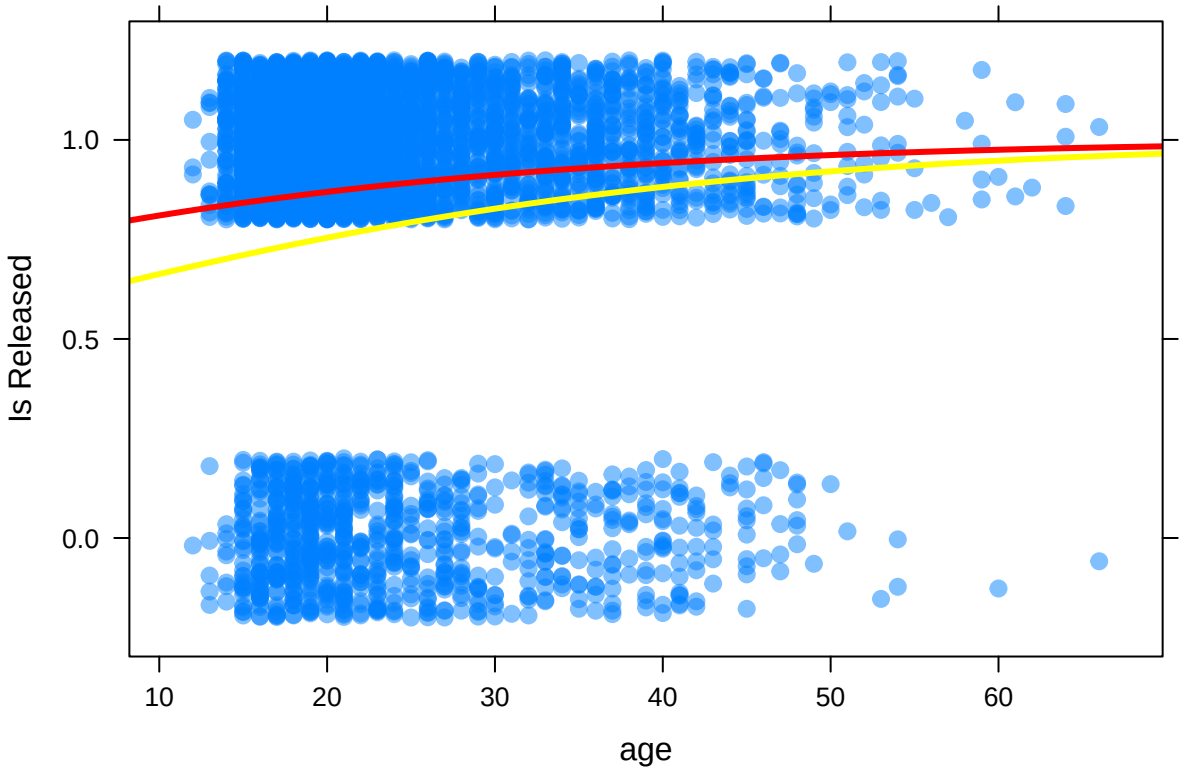
XYplot



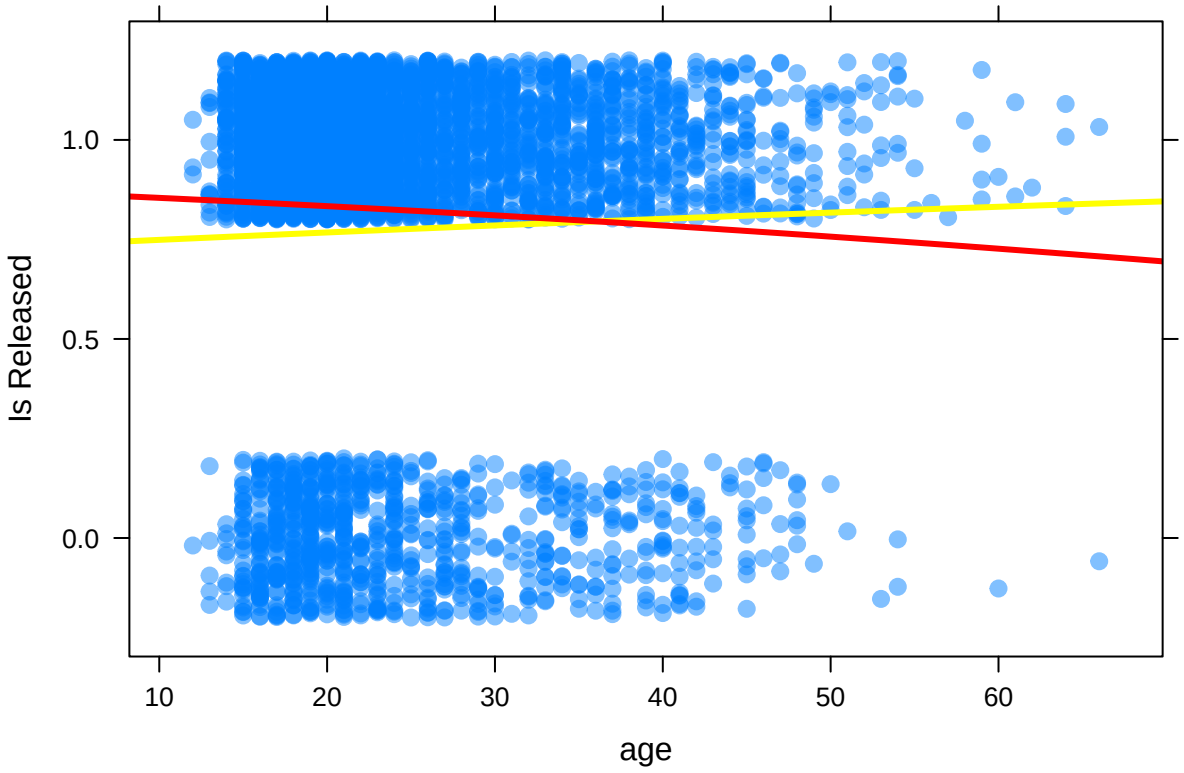
Based on the best logistic model, now we use xyplot to investigate two main interested questions? 1). whether employed is helpful for improving probability of releaseing after arrested? 2). whether being a citizen is helpful for improving probability of releaseing after arrested? Here, we take the most recent year 2002, and for a Black male who is firstly arrested(checks = 0) to investigate the questions using xyplots firstly and then we do the same for a white male with same conditions, we do it seperately for white and black people as there seems some differences between white and black in probability of released as discrination of race existed:



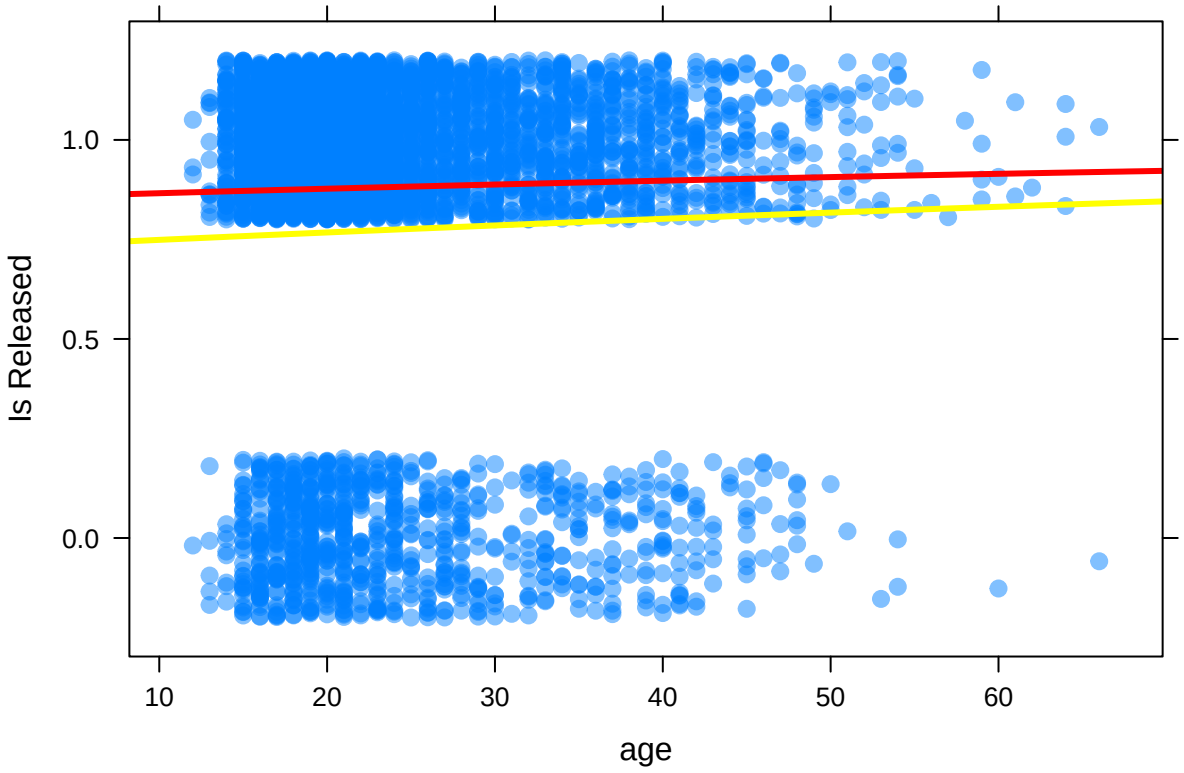
From the above xyplot, given the people is no citizen, it can be founded that for youngers(age < 35), being employed has a much higher released probability than those not being employed as the redline(employed) is much higher the yellowline(not employed) in the xyplot, and they are close for old people as for old people, most of them are already not employed(retired).



From the above xyplot, given the people is not employed, it can be founded that being citizen has a much higher released probability than those not being citizen as the redline(citizen) is much higher the yellow-line(not citizen) in the xyplot.



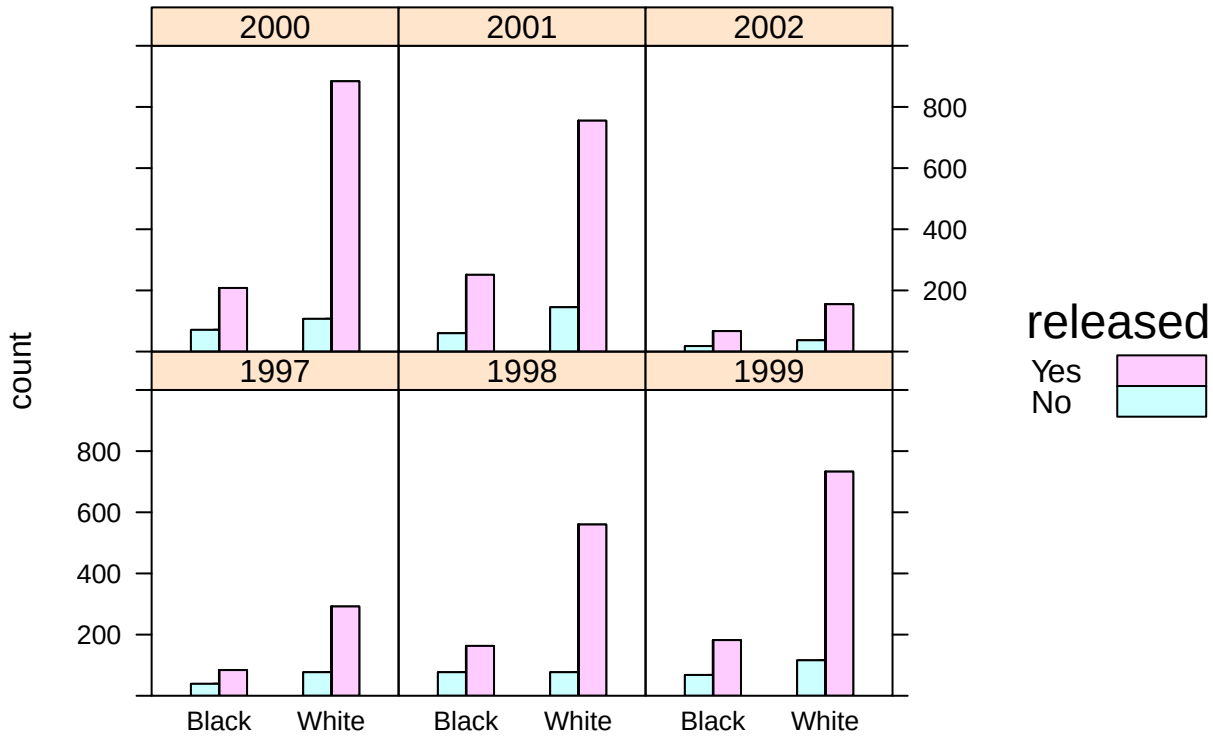
From the above xyplot, given the people is no citizen, it can be founded that for youngers, being employed has a much higher released probability than those not being employed as the redline(employed) is much higher the yellowline(not employed) in the xyplot, but for olders, being employed has a little lower released probability than those not being employed as the redline(employed) is little lower than the yellowline(not employed).

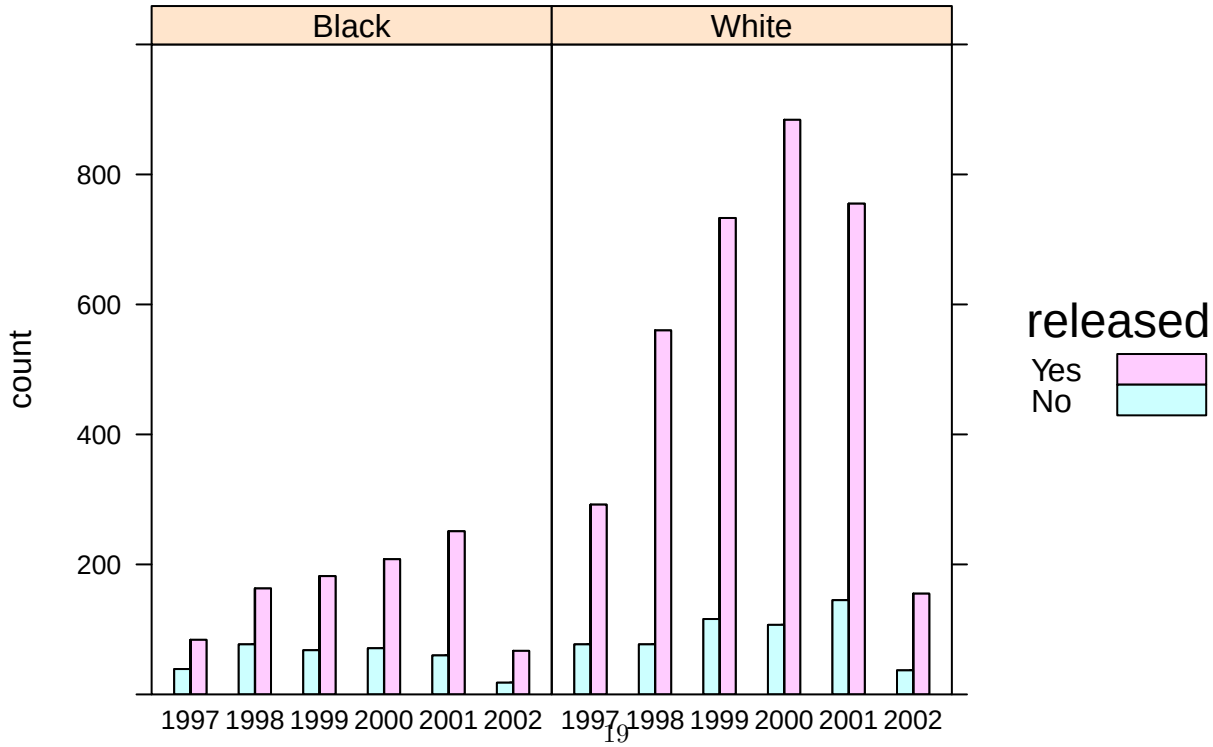


From the above xyplot, given the people is not employed, it can be founded that being citizen has a much higher released probability than those not being citizen as the redline(citizen) is much higher the yellowline(not citizen) in the xyplot. So that, overall, we can conclude that being employed and being citizen could help improving probability of releaseing after arrested a lot for both white and black young people.

Frequency bar charts

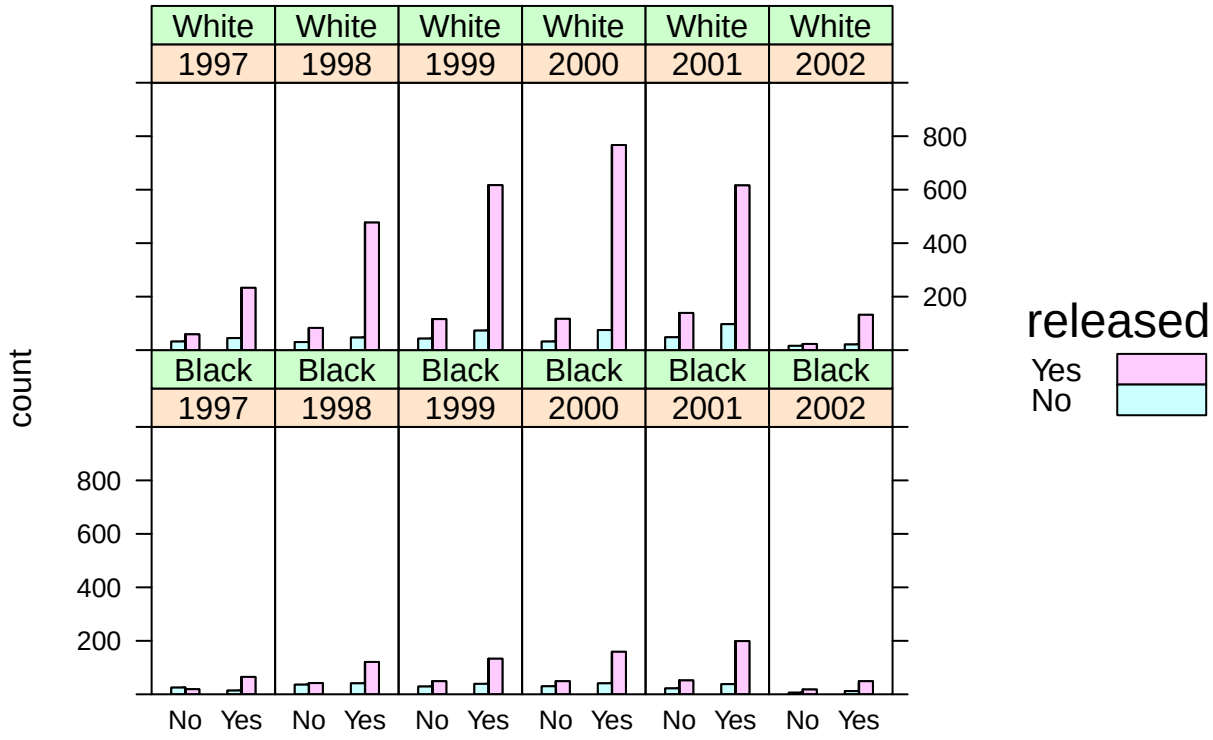
The Relationship between 'colour' and 'released' as year changes





As we mentioned before, the interaction between colour and year is significant. So we explored the relationship between colour and the response variable in different years. As shown in the graph, the release rate among the white is always higher than among the black. Thus there always exists racial discrimination from 1997 to 2002. And the gap of release rate between the white and the black gets larger from 1997 to 2000, but it gets smaller from 2000 to 2002. Therefore we can see the discrimination of race is most serious in the year of 2000, and it is least severe in 2002. The situation of discrimination differs in different years.

The relationship between 'employed' and 'released'



The interaction term of employed and year is significant in our model, and since colour is a potential confounding factor, so we when we look at the influence of employed on release rate in different years , we need to control the 'colour' factor. From the bar chart , we can see that when we control the colour to be white, the gap of the release rate among those who are employed are overall higher than those who are not employed. And the gap of the release rate becomes larger and larger from year 1997 to 2000, and becomes smaller from 2000 to 2002. ANd for the black group, in the year of 1997, the trend is similar to the white group, only slightly different, the gap of the release rate is larger from 2017 to 2001, and the gap comes to a peak in year 2001, then the gap shrinks in the year 2002. In other words, the discrimination with respect to employment varies in different years. For the white people, discrimination is most severe in the year 2000, and least severe in 2002. And for the black people, discrimination is most severe in 2001, and least severe in 2002 if we look at the gap between the release rate among different groups.

Conclusion

For model comparison part, we can see the best model also performs well in prediction evaluation. And in the added variable plot we can see that the interaction terms based on 'colour' have a strong influence on

the dependent variable 'released'. From the xy plot, we can conclude that being employed and being citizen could increase the probability of being released greatly for both white and black young people. As we see that 'colour' is a potential confounding factor, so we separate them into two groups, which are the white and the black. By looking at the frequency plots for colour-released we see that there exists racial discrimination from 1997 to 2002 since overall the release rate for white people is higher than the black. The situation of racial discrimination varies in different years, and the situation is the most serious in 2000. For the employed -released plot, it is clear to see that the release rate of people who are employed are higher than those who are not employed. Thus the discrimination with respect to employment always exists from 1997 to 2002, but also varies in different years when we control the factor 'colour'.