Proceedings of the 2nd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/17/20231066

Whether a Criminal is Likely to Re-offend? A Statistical Analysis Using Crimes Data from Broward County Florida

Xinyuan Wang[1,a,*], Mingxin Liu[2], Chengran Song[3], Hongru Tan[4]

[1]*Faculty of Art and Science, University of Toronto, Toronto, M5S 1A4, Canada*
[2]*The Affiliated High School of Peking University; Beijing;100086; China*
[3]*College of Arts and Science; Boston University; Boston;02215; America*
[4]*Shenzhen college of international education; Shenzhen, 518000, China*
a. vivalagloria.wang@mail.utoronto.ca
**corresponding author*

Abstract: Security issues have always been a significant threat to the safety of citizens in every country, and many of these re-arrested criminals have negatively impacted social security. Therefore, predicting and studying the factors of a criminal's re-entry to prison will significantly help maintain social order and improve the civil society happiness index. This study, it will show what elements are predicted to influence a criminal's return to prison and what aspects will have a higher proportion and weight based on the collected data set. In the dataset, each re-admission inmate is categorized according to gender, age range, race, records, etc. Use the Logistics model and OLS model to build a model to predict what factors most directly lead to a criminal being arrested and imprisoned again. Data research has proved that the "number of priors" is the factor that most affects the recidivism rate of criminals.

Keywords: statistics, prediction of recidivism rate, ordinary least-squared model, logistic regression model, social inequality

1. Introduction

The problem of criminals has always been the most severe social security problem in most countries. In many countries, criminals continue to disrupt public safety, so there must be some commonality among these criminals. First of all, many criminals are criminals with previous convictions, so this part is the most research oriented. So, we're going to look at what factors make some criminals more likely to re-offend and go to jail. Before the analysis, we listed some factors that may affect the crime, including the sex of the offender, the age range of the offender, the criminal record of the offender, etc. Among these factors, we defined gender as male and female, age group into three, and criminal record as yes or no. Next, we give weights to different factors and parts and use the logistic regression model to predict. In such a prediction process, it can be found that the X factor will be the most influential factor for a criminal to be caught again, and Y and Z also have a particular influence on the criminal's re-offending.

DOI: 10.54254/2754-1169/17/20231066

2. Literature Review

Predicting whether a criminal will re-offend is a highly discussed topic, and many researchers have addressed the content of this topic. At Bruce Frederick, Ph.D. "Factors Contributing to Recidivism Among Youth Placed with The New York State Division For Youth" [1].

A similar study has been published in the article. In his research, he focused on what causes teens in the New York area to be imprisoned again and, at the same time, predicted their first recidivism. In such a project, Bruce Frederick first used the data of 9,477 repeat offenders. Based on their situation, he proposed the gender, race, history of alcohol and drug abuse, and whether they had received good aftercare after being released from prison., whether you have received an education, etc., as a factor to consider. At the same time, Frederick also thought all the previous crimes' recidivism. He divided them into four categories according to the reason for entering the prison and the time of detention. During the analysis process, Federick talked about the data he used and conducted a lot of chart and model analysis, among which he used a lot of bar charts and table charts to explain every single item and increase the contrast of conditions. After the establishment and analysis of the model, Frederick first thought that males would be a characteristic of the recidivist population, and secondly, whether the community environment of the recidivist was an environment conducive to the physical and mental growth of these juveniles. Whether the enlightenment and education work is completed seriously is also an important factor affecting their re-offending. Frederick's research details recidivism factors that affect youth in the New York area. The same theme is reflected in the article "The Dangers of Risk Prediction in the Criminal Justice System" by Julia Dressel and Hany Farid [2]. But the difference is that Dressel and Farid's article focuses on bias and fairness in predicting whether a criminal will re-offend. In addition, research stresses the need for court actors to consider multiple aspects of the criminal record. However, most quantitative studies of sentencing have found that the number of previous convictions or arrests affects the criminal record [3]. Other studies have shown that rehabilitation is effective in reducing recidivism. But the offender's sex, age, or level of addiction can affect recidivism [4].

This will also be a topical topic because there will be racism in the cognition of some people, so in essence, these people will think that blacks are more recidivists than whites. But the actual situation is not the case. After a large amount of data calculation and model analysis, it is confirmed that the reason for the recidivism of criminals does not have any direct relationship with people. This article made me understand and realize in great detail that any subjective speculation in the process of creating an analysis and prediction is ridiculous and contradictory to reality. Therefore, we drank 6172 data in fact in detail in our research topic and simultaneously proposed several possible reasons, using the modeling and charting to analyze and integrate the data.

3. Methodology

The logistic Regression Model is widely used for classification tasks. The model predicts the probability an object belongs to a particular group; In other words, the chances this object has a specific feature or not. **YES** is written as **1** and **NO** is written as **0** in the dataset.

Step 1: Linear regression

To begin with, we need a regression line. This line comes from linear regression.

DOI: 10.54254/2754-1169/17/20231066

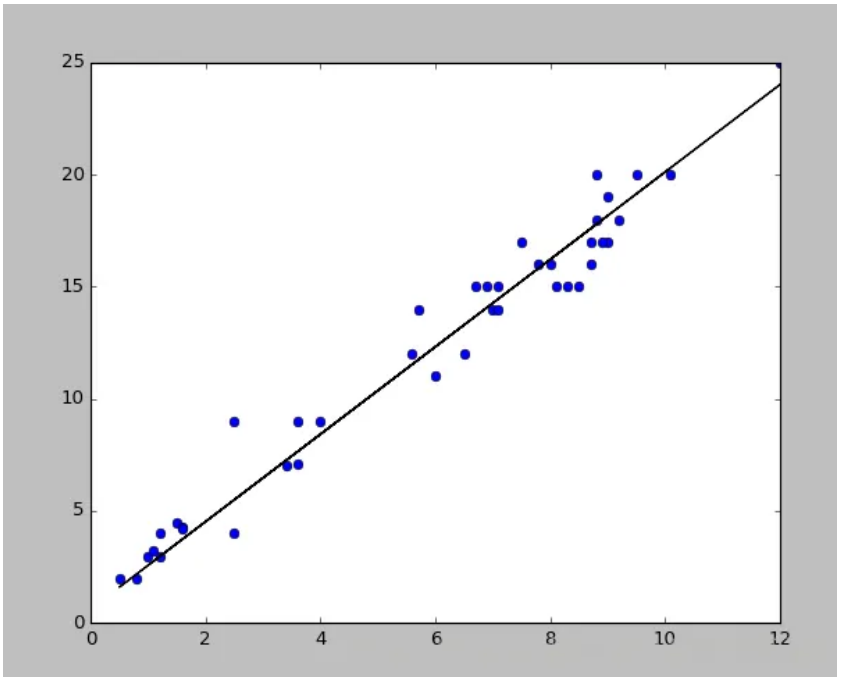

Figure 1: A common example of the linear regression line.

In this case, the line could be:

$$Y\hat{} = \beta 0 + \beta 1X1 + \beta 2X2$$

Step 2: Sigmoid function [5]
Sigmoid function is written as:

$$S(x) = \frac{1}{1 - e^{-x}}$$

S(x) value will always be between 0 and 1, no matter the x values. The graph of sigmoid function is:

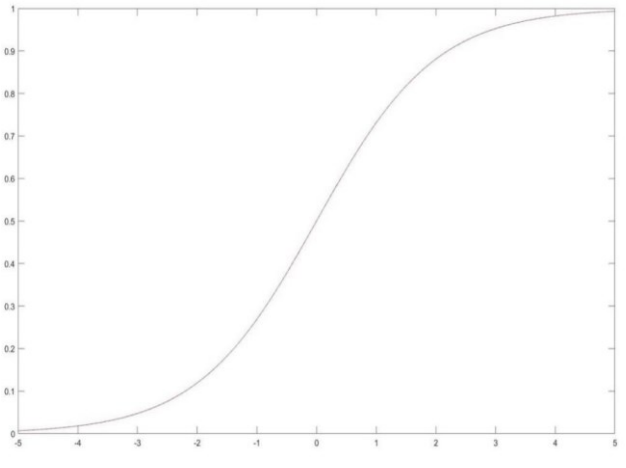

Figure 2:The graph of a sigmoid(x) function.

As shown from Fig.2., the function has two asymptotes: 1 when x approaches positive infinity; 0

DOI: 10.54254/2754-1169/17/20231066

when x approaches negative infinity. The middle line of the function is x = 0, and y equals to 0.5. Usually in a logistic regression model, 0.5 is the threshold. S(x) values greater than 0.5 are considered as 1; S(x) values smaller than 0.5 are considered as 0. Hence, S(x) value is said to be the probability that S(x) belongs to 1 or 0. A S(x) values of 0.788 is considered as 1, and the probability that 0.788 equals to 1 is 78.8%.

Step 3: Application

Assuming we want to investigate whether a tumor is benign or malignant, we collect data from 1000 patients who have been through tumors before. X variables will be several indexes such as whether they are overweighted, alcoholic, young or old. Y values are either 1 or 0. (1 if the tumor is malignant, 0 if the tumor is benign). Linear regression is carried out in the data, and a regression line is obtained. Next, 1000x values are put in regression line and 1000y values are obtained. In this case, y values obtained will be the x values of sigmoid function(S(x)). 1000 S(x) values are obtained. If S(x) is greater than 0.5, label the x (y value for the regression line) as 1; If S(x) is smaller than 0.5, label the x as 0. Finally, compare the predicted 1 and 0 values with the Y values of raw data. If the precision rate is higher than 90%, the model is considered as a good model and can be used to predict the nature of a new tumor.

4. Data Analysis

4.1. Data Acquisition and Preprocessing

We use the excel data containing a sample of 6,172 individuals illustrating their basic personal information, including gender, race, age, whether they conducted crimes before, and whether the previous case is a misdemeanor or a severe one.

In Table 1, shown below, the data is all categorized as binary data given 6,172 observations, as the minimum and maximum values are always 0 and 1. Reid is an abbreviation of recidivism and is labeled as 1 if the individual we are investigating did conduct a crime again within two years; old equals 1 infers that if the age of the individual is above forty-five and young refers to those aged below twenty-five. Other variables briefly all indicate the ethnicity of the chosen individual, and if they qualified as one of the races, they are labeled as 1 under that given variable.

Table 1: OLS model summary of a given dataset.

Variable	Obs	Mean	Min	Max
recid	6,172	0.455	0	1
male	6,172	0.810	0	1
old	6,172	0.209	0	1
misdem	6,172	0.357	0	1
young	6,172	0.218	0	1
af_am	6,172	0.514	0	1
caucasian	6,172	0.341	0	1
hisp	6,172	0.082	0	1
asian	6,172	0.005	0	1
nat_am	6,172	0.002	0	1

Since we are trying to predict the possibility of whether a criminal is likely to re-offend, we introduce ordinary-least squared data analysis as it shows a direct way to establish the relationship between individual variables and the given dependent variable. As noted above, the data follows

DOI: 10.54254/2754-1169/17/20231066

binary distribution; hence logistic model can also be brought up under our further analysis.

4.2. The OLS Multivariate Linear Regression Analysis

At first, as shown in Table 2, we use an ordinary least-square linear regression model to briefly illustrate the overall influence of each variable without making any changes to it [6]. We notice that with a 95% confidence interval, we need the p-value to be less than 0.05 to satisfy the condition that those predictor variables without manipulation are statistically significant towards recidivism, and from the figure below, most of the variables fit this condition. However, the p-values of the ethnicity variables, including caucasian, hisp, native American, and Asian, exceed the 5% significance level, which indicates they are not a relatively significant variable in explaining recidivism.

Table 2: OLS model numerical data result of relationship among originally given variables.

recid	Coef.	Std.Err.	t	P>|t|
Male	0.124	0.016	7.84	0.000
old	-0.127	0.016	-7.98	0.000
young	0.087	0.016	5.61	0.000
af_am	0.150	0.028	5.41	0.000
caucasian	0.051	0.028	1.81	0.070
hisp	0.008	0.034	0.25	0.805
nat_am	0.093	0.149	0.62	0.533
asian	-0.097	0.091	-1.07	0.287
_cons	0.268	0.030	8.98	0.000

Next, we want to add more possible influential variables into the OLS model, as shown in Tables 3 and 1) firstly, we focus on the effect of age and gender working together on recidivism. We see that the coefficient of the probability of being male itself causing re-offending is approximately 0.124 from Table 2 above. Still, when we combine the effect of being male and being young, the state result shows that the coefficient of the possibility of being male decreased to 0.097 instead; being male and young at the same has a p-value of 0.005 shown in step 1 column. Hence, we can conclude that being male and young simultaneously is an important factor in recidivism. 2) Moving towards the next step, we analyze whether ethnicity itself works as a significant independent factor or whether recidivism is impacted by gender and ethnicity together. Similarly from above, if an individual is an African American, he increases the possible rate of re-offending by 0.15, but if we add in the effect of being male and African American at the same time, it shows that the effect of being African American itself drops to 0.042 with a not statistically significant p-value equalling to 0.248 which says it's not the ethnicity that leads to recidivism, but gender also plays a significant role here. Being male and African American at the same time seems to be a relatively notable factor counting towards recidivism. 3) Thirdly, we add the variable male_misdemeanor into our model, it shows a similar process as the previous two steps, and we find a similar result that if a male individual conducted a comparatively small crime or mistake before, it also reflects a higher rate of re-offending [3]. The p-value of the independent variable male_misdemeanor is 0.045, which is regarded as statistically essential but not that unignorable.

DOI: 10.54254/2754-1169/17/20231066

Table 3: OLS model numerical data results of different new-defined variables.

recid	**Coef.(1)**	**P>|t|**	**Coef.(2)**	**P>|t|**
male	0.097	0.000	0.038	0.086
old	-0.109	0.001	-0.123	0.000
young	-0.011	0.760	0.001	0.984
misdem	-0.093	0.000	-0.095	0.000
af_am	0.143	0.000	0.042	0.248
caucasian	0.052	0.056	0.049	0.074
hisp	0.012	0.719	0.012	0.704
nat_am	0.086	0.583	0.087	0.571
asian	-0.098	0.234	-0.092	0.264
male_young	0.112	0.005	0.099	0.010
male_old	-0.017	0.655	-	-
male_af_am	-	-	0.123	0.000
male_misdem	-	-	-	-
_cons	0.326	0.000	0.375	0.000

Table 4: Coefficients of OLS model.

recid	Coef.(3)	P>|t|
male	0.009	0.754
old	-0.123	0.000
young	0.001	0.983
misdem	-0.144	0.000
af_am	0.036	0.331
caucasian	0.049	0.072
hisp	0.013	0.688
nat_am	0.082	0.588
asian	-0.093	0.258
male_young	0.100	0.009
male_old	-	-
male_af_am	0.130	0.000
male_misdem	0.063	0.045
_cons	0.399	0.000

Throught out the process, based on Table 3, the final OLS model we generated is

Recidivism= 0.009 (male) + -0.123(old) + 0.001(young) + -0.144(misdemeanor) + 0.036(af_am) + 0.049(caucasian) + 0.013(hisp)+ 0.082(nat_am) + -0.093(asian) + 0.100(male_young) + 0.130(male_af_am) + 0.063(male_misdemeanor) as we add in those variables we considered as important.

Theoretical Reasons Why the OLS Model isn't a Good Model:

To be confident about the multivariate linear regression model we conducted, there are some classical assumptions we need to consider: 1. homoskedasticity: the variance of the residuals has to be homoscedastic or constant to make sure we have precise standard error. 2. multicollinearity: the relationship between predictor variables is not perfectly multicolinear.3. normality: residuals are normally distributed.

Homoskedasticity.

DOI: 10.54254/2754-1169/17/20231066

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of recid

         chi2(1)      =    12.58
         Prob > chi2  =   0.0004
```

Figure 3: Homoskedasticity test result of defined OLS model.

As shown in Figure 2, The Breusch-Pagan test results at the 95% confidence level show that the p-value equal to 0.0004 is less than 0.05; hence we reject the null hypothesis Ho: constant variance; thus, the residuals do not display constant variance. Our current model fails the homoskedasticity test.

Multicollinearity.

Table 5: VIF test result of multicollinearity of defined OLS model.

Variable	**VIF**	**1/VIF**
af_am	9.63	0.10
male_af_am	6.59	0.15
male_young	5.80	0.17
young	5.57	0.18
male_misdem	5.53	0.18
misdem	5.06	0.20
caucasian	4.72	0.21
male	3.03	0.33
hisp	2.28	0.44
old	1.11	0.90
asian	1.09	0.92
nat_am	1.03	0.97
Mean VIF	4.29	-

Table 5 shows that the variables we analyzed all have VIF < 10, indicating the relationship among independent variables is not perfectly multicollinear. Our model passes the multicollinearity test.

Omitted Variable Test.

```
Ramsey RESET test using powers of the fitted values of recid
      Ho:  model has no omitted variables
               F(3, 6156) =      1.60
                 Prob > F =      0.1864
```

Figure 4: Omitted variable test result of defined OLS model.

DOI: 10.54254/2754-1169/17/20231066

The null hypothesis is that the model does not have omitted-variables bias and as the p-value is higher than the usual threshold of 0.05 (95% significance), we fail to reject the null hypothesis and conclude that we do not need more variables [7]. However, we still need to consider the possibility of other independent variables influencing recidivism, including educational background, social status, etc. The problem is that those omitted factors could be indirectly related to the ones we are currently analyzing, making us fail to reject the null hypothesis.

Normality.

Our model failed the normality test as the data follows the binary distribution.

But the normality test helps us introduce the logistic regression model as the data distribution shows approximately a binomial distribution which is precisely how the logistic regression model works. Even though we could add more variables that might affect recidivism based on the current data we had concerning different predicators, there are plenty of problems within the OLS model itself. Based on those points, we wouldn't consider using OLS model to analyze our data. Hence, we introduce the logistic model.

4.3. Logistic Regression Model

The logistic regression model is used because the above OLS model cannot directly predict or obtain the probability of an individual offender reoffending. The logistic regression model is suitable because it can fit the binary variables "Yes" or "No" perfectly. In addition, it can also derive the predicted probability of reoffending from the training set. R language was used to generate the logistic regression model in the research.

4.3.1. Single Variable: Basics and Optimization

In Figure 4 shown below, the number of stars behind the p-value means the significance.

Only a tiny p-value here refers to the tremendous significance.

The p-value of Caucasians and Hispanics is equal to 0.12 and 0.905858. It is clear now that Caucasian and Hispanic people, these two explanatory variables have no significant effects on the dependent variable "recidivism". That also highly corresponded to finding in the ordinary linear model.

Deleting these two variables without significant effects on the response variable can optimize the model and let us get a more ideal one.

```
> recidivism <- glm(recid ~ number_of_priors + + misdem + af_am + young + old + caucasian
+ hisp+male, data=recidivism_data_for_project, family=binomial())
> summary(recidivism)

Call:
glm(formula = recid ~ number_of_priors + +misdem + af_am + young +
    old + caucasian + hisp + male, family = binomial(), data = recidivism_data_for_projec
t)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7431  -0.9690  -0.6463   1.0857   2.0686

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.60641    0.13235  -4.582 4.61e-06 ***
number_of_priors  0.78533    0.03825  20.532  < 2e-16 ***
misdem           -0.21872    0.05875  -3.723 0.000197 ***
af_am             0.28386    0.11928   2.380 0.017325 *
young             0.73328    0.06893  10.638  < 2e-16 ***
old              -0.66945    0.07607  -8.800  < 2e-16 ***
caucasian         0.18798    0.12195   1.541 0.123219
hisp              0.01758    0.14864   0.118 0.905858
male              0.34771    0.07185   4.840 1.30e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8506.4  on 6171  degrees of freedom
Residual deviance: 7577.0  on 6163  degrees of freedom
AIC: 7595

Number of Fisher Scoring iterations: 4
```

Figure 5: Summary of initial logistic model in R, single variable.

DOI: 10.54254/2754-1169/17/20231066

```
> recidivism <- glm(recid ~ number_of_priors + misdem + af_am + young + old + male, data=
recidivism_data_for_project, family=binomial())
> summary(recidivism)

Call:
glm(formula = recid ~ number_of_priors + misdem + af_am + young +
    old + male, family = binomial(), data = recidivism_data_for_project)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7456  -0.9714  -0.6652   1.0875   2.0129

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.46518    0.07857  -5.921 3.21e-09 ***
number_of_priors  0.78713    0.03828  20.563  < 2e-16 ***
misdem          -0.21938    0.05871  -3.736 0.000187 ***
af_am            0.14939    0.05694   2.624 0.008700 **
young            0.73028    0.06887  10.604  < 2e-16 ***
old             -0.66132    0.07595  -8.708  < 2e-16 ***
male             0.33946    0.07171   4.734 2.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8506.4  on 6171  degrees of freedom
Residual deviance: 7581.2  on 6165  degrees of freedom
AIC: 7595.2

Number of Fisher Scoring iterations: 4
```

Figure 6: Summary of optimized logistic model in R, single variable.

As shown in Figure 6, the function of predicting recidivism probability is:

$$\text{recid} = -0.66132(\text{old}) + 0.73028(\text{young}) + 0.78713(\text{priors}) + -0.21938(\text{misdem}) + 0.14939(\text{af\_am}) + 0.33946(\text{male}).$$

The function of the logistics model with a single variable is generally similar to the part of OLS model.

Table 6: Confident interval of logistic model, single variable.

	2.5%	97.5%
Intercept	-0.62	-0.31
Number_of_priors	0.71	0.86
misdem	-0.33	-0.10
af_am	0.037	0.261
young	0.60	0.866
old	-0.81	-0.51
male	0.199	0.481

Table 7: Coefficients logistic model, single variable.

Name	Intercept	Number_of_priors	misdem	Af_am	Young	Old	male
Estimate	-0.47	0.79	-0.219	0.15	0.73	-0.661	0.34
P_value	3.21e-9	<2e-16	0.00019	0.0087	<2e-16	<2e-16	2.21e-6

As we can see from Table 6 and 7, the 95% confidence intervals of each factor, each x variable, do not include 0, indicating that each element will affect recidivism. The P-value is all small enough (p-value < 0.05). In this case, this model is optimized. It's safe for us to use.

DOI: 10.54254/2754-1169/17/20231066

4.3.2. Regression Diagnostics

When the response variable has a limited number of values (such as logistic regression, it only has the binary variables in the dataset, “Yes” or “No”), the power of the diagnostic map is much reduced. Normality can’t be checked in the logistics model because they are not linear. In this case, partial approaches are used.

1.Linear Checking

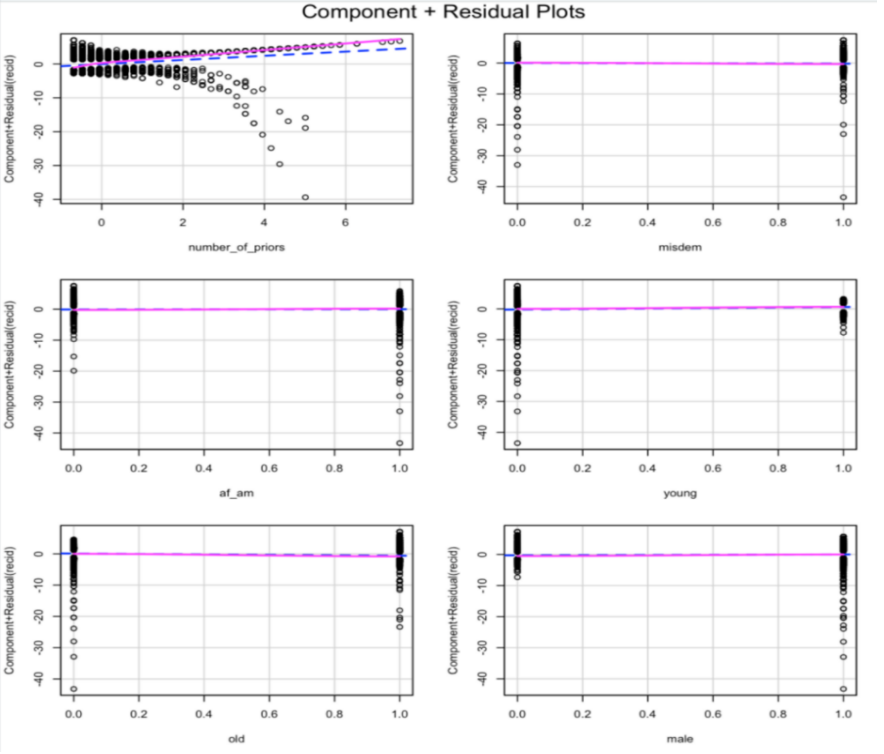

Figure 7: Linear checking.

In Figure 7, each fitted lines are close to horizontal. There is a solid linear relationship between reoffending and each independent variable.

2.Detection of Multicollinearity

```
> vif(recidivism)
number_of_priors          misdem           af_am           young
        1.140197        1.027530        1.065453        1.151107
             old            male
        1.093692        1.009400
> sqrt(vif(recidivism)) > 2 # problem?
number_of_priors          misdem           af_am           young
           FALSE           FALSE           FALSE           FALSE
             old            male
           FALSE           FALSE
```

Figure 8: Detection of multicollinearity.

They are using the variance inflation factor to check the errors of the model. The sign of the evo-

DOI: 10.54254/2754-1169/17/20231066

lution of their vif value is all lower than 2, indicating no multicollinearity problem in the logistics model.

3.Outlier Test

```
> outlierTest(recidivism)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
5541 -2.752534         0.0059136            NA
```

Figure 9: Detection of outlier.

This function in R module determines whether there are outliers based on the significance of a single maximum (positive or negative) residual. P=0.0059136, more diminutive than 0.05, demonstrates no outliers in the data set.

4.Influence Points

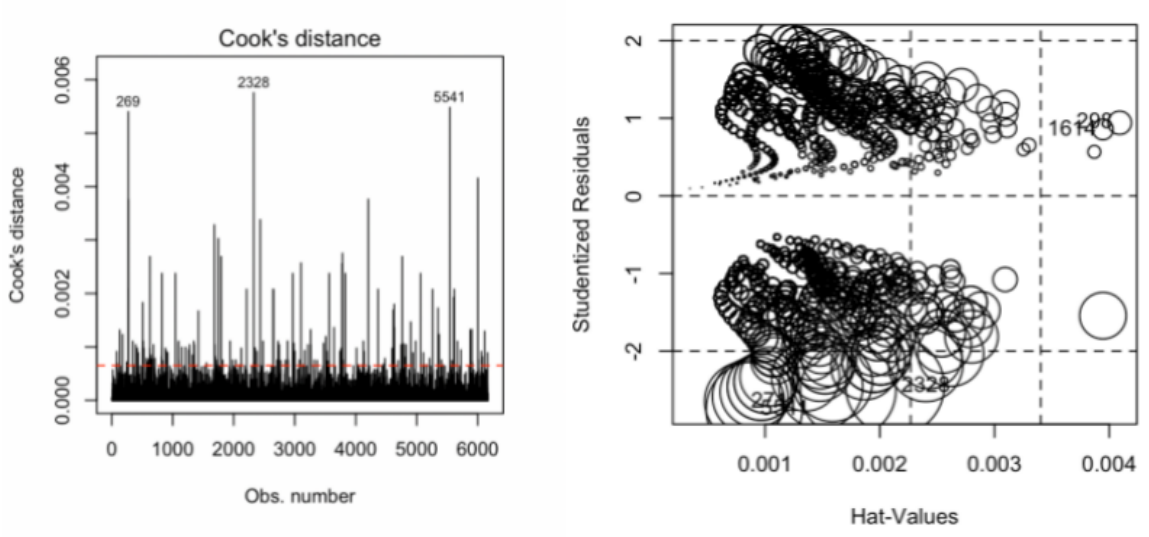

Figure 10: Influence points.

There are three influence points, 269,2328, and 5541. If one of them changes its numerical size, the pattern of the entire data set will be dramatically changed. Their leading group is nearly all within between absolute value 2 of studentized residuals, which is reasonable, and made diagnosis is relatively successful.

4.3.3. Training Data Set

From Table 8, on the contrary, for those older than 25, the probability of reoffending is multiplied by 0.5161718.

DOI: 10.54254/2754-1169/17/20231066

Table 8: Exponentiate coefficients, single variable.

	Exponentiate Coefficient of Recidivism
Intercept	0.63
Number_of_priors	2.20
Misdemeanor	0.80
Af_am	1.16
Young	2.08
Old	0.52
Male	1.40

Having the features that a criminal record, af_am, young, male parts appear in a criminal, the probability of reoffending will increase. However, the misdemeanor and old(age above 25) features appear on the offender, and the likelihood of his reoffense decreases, like the "old" variable, nearly smaller than its original two times. These factors or traits can be present in a person simultaneously, except old and young are mutually exclusive.

The predictor variable cannot be equal to 0; the intercept term has no particular meaning here.

Table 9: Prediction of recidivism probability.

	old	misdem	priors	af_am	young	male
old	1	---	---	---	---	---
misdem	---	1	---	---	---	---
priors	---	---	1	---	---	---
af_am	---	---	---	1	---	---
young	---	---	---	---	1	---
male	---	---	---	---	---	1
Tested Probability	0.33	0.42	0.65	0.48	0.60	0.47

From Table 9, it can be intuitively found that when the occurrence of each feature is 1, the recidivism probability is: number of priors > young > af_am > male > misdem > old.

The most significant factor that increases the probability of re-offending is the number of criminal priors, roughly 0.65. The more priors, the more likely the criminal is to re-offend. The second most crucial influencer is young, followed by African Americans, male, misdemeanor, and finally old, 0.3321.

The conclusion of the test dataset successfully validates the functionality of the model. Furthermore, the example of the training set is shown in Figure 11. Only the number of prior increased steadily, and all other values are controlled as its averages and unchanged.

DOI: 10.54254/2754-1169/17/20231066

Testdata_priors

Name	Af_am	Num_of_p	misdem	young	Old	male	prob
1	--	1	--	--	--	--	0.649
2	--	2	--	--	--	--	0.803
3	--	3	--	--	--	--	0.899
4	--	4	--	--	--	--	0.952
5	--	5	--	--	--	--	0.977

Figure 11: Predication of recidivism probability of "Number of Priors".

4.3.4. Interaction Variable: Model Basic Information

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.014720   0.125958  -0.117 0.906970
number_of_priors         1.199087   0.133271   8.997  < 2e-16 ***
af_am                   -0.262878   0.138182  -1.902 0.057118 .
young                    0.676515   0.190697   3.548 0.000389 ***
old                     -0.689000   0.192524  -3.579 0.000345 ***
male                    -0.173375   0.137675  -1.259 0.207921
misdem                  -0.380012   0.145021  -2.620 0.008783 **
young:male               0.456287   0.183435   2.487 0.012866 *
old:male                 0.066511   0.202503   0.328 0.742574
af_am:male               0.475552   0.150910   3.151 0.001626 **
male:misdem              0.230766   0.154945   1.489 0.136398
old:misdem              -0.002093   0.150445  -0.014 0.988899
number_of_priors:male   -0.274332   0.125420  -2.187 0.028720 *
number_of_priors:af_am  -0.216494   0.078870  -2.745 0.006052 **
number_of_priors:young   0.666613   0.197846   3.369 0.000753 ***
number_of_priors:old    -0.193219   0.077150  -2.504 0.012264 *
number_of_priors:misdem  0.047953   0.080905   0.593 0.553371
```

Figure 12: Summary of logistic model, interaction variable.

Concluding from Figure 11, the influence of many parameters on "recidivism" has changed with the addition of different interaction terms. However, the overall pattern has changed little.

Table 10: Exponentiate coefficient, interaction variable.

X Variables	Exponentiate Coefficients
priors	2.63
Young (age <25)	1.83
Old (age > 25)	0.51
Misdemeanor	0.65
Young* male	1.73
Af_am* male	1.52
Priors* af_am	0.80
Priors* young	1.97
Priors* old	0.82

DOI: 10.54254/2754-1169/17/20231066

In Table 10, it can take "prior*young" as an example. The interpretation for this chart can be if this crime has a prior conviction, they are young (the age is above 25), their recidivism probability will multiply 1.97.

1. It's similar to prior single variable logistics regression. The most significant variable is still the "number of prior convictions". The recidivism will be multiplied by 2.63. We can safely believe that if this crime has a conviction history or is recorded, he has the most significant probability of recidivism, contrasting with other features. This is the specific group that police have to pay more focus to. Similarly, older offenders are the least likely to commit again since their reoffending probability is nearly negligible twice, multiplied by 0.51.

2. Interaction variables are more convincing than single variables. Consequently, "prior*young" is the most significant explanatory variable affecting recidivism. It can say that the most likely offenders have the "criminal record and young" trait, and the probability of reoffending will be multiplied by 1.97. When "prior of conviction" meets the "old" or "af_am" trait, they will reduce the likelihood of reoffending caused by "prior of conviction", since they are predicting recidivism probability would be multiplied by 0.82 and 0.79.

3. The concluded formula:

$$recid = 2.6272(priors) + 1.8320(young) + 0.5094(old) + 0.6490(misdeem) + 1.73(young * male) + 1.52(af\_am * male) + 0.80(priors * af\_am) + 1.97(priors * young) + 0.8171(priors * old)$$

This is more honorable than the one in the single variable logistics model.

4. Racial Bias

af_am	-0.262878	0.138182	-1.902	0.057118
caucasian	0.18798	0.12195	1.541	0.123219
hisp	0.01758	0.14864	0.118	0.905858

Figure 13: The contrast between af_am with other terms.

Figure 12, it shows that racial bias is not apparent. This is because if the variables are compared, af_am has a more negligible recidivism probability than other characteristics such as "prior number" or "youth". "af_am" isn't even significant because its p-value is too large. The pattern for African Americans is similar to that for Caucasian and Hispanic.

According to what is often said in society, African Americans always commit crimes, which is a false prejudice. In terms of the model, it is not significantly biased against African Americans.

4.4. Discussion

The conclusions from the logistic regression model are credible.

The variable with the most influence is the "number of priors". The criminal's constant crime shows that his living circumstances force him. One possible reason could be speculated to be poli-

DOI: 10.54254/2754-1169/17/20231066

cy. The re-offender may be influenced by social inequality, such as racial prejudice and gender inequality. After they left prison, they perceived "discrimination against criminals" by society [8]. As a result, those offenders can not find regular jobs. In this case, the offender had to turn to crime to support his livelihood. They are kind of unequal social groups. It is accessible to do some mix-method research to solve the problem in the future.

It makes sense that older criminals will re-offend less. As criminals get older, their physical fitness declines. It gets in the way of their crimes. In addition, these criminals set up their families, even take on responsibilities for caring for children or become more stable, all potential causes of lower recidivism among older offenders.

5. Limitation

5.1. The Limitation of the OLS Model

1. With R-squared being 0.0615 for the final OLS model, which means it covers only approximately 6 percent variances of the variables, it is not a convincing result to use.

2. It doesn't directly show the probability of whether an individual might re-offend, given other essential factors. In other words, the OLS model indirectly reflects the result we seek.

3. The OLS model failed some classical assumptions of being a suitable or acceptable good model.

5.2. The Limitation of the Logistics Model

1. Although scholars who support systems that predict criminal procedures argue that big data and advanced machine learning make these analyses more accurate and less biased than humans [9]. However, the main limitation of logistic regression is the assumption of linearity between dependent and independent variables [10]. Nonlinear r problems cannot be solved by logistic regression. Interaction variables must be used in many places in our data set, which may lead to inaccurate data.

2. In the future, we can expand the dataset. For example, We can collect more independent variables, like observing more traits. We can also use more diversified methods, such as decision trees.

6. Conclusion

Through Logistics and OLS models, the "number of priors" is the factor that most affects the recidivism rate of criminals. It has contributed to the increase in recidivism more than any other factor in this data set. On the contrary, "old" is the minor factor affecting the recidivism rate of criminals in the data set. The older people are, the lower the recidivism rate. In our data analysis, "Hispanic", "African_American" and "Caucasian" are not counted as influencing factors because the hypothesis test does not have enough evidence of them (p-value is too small).

Acknowledgment

All the authors contributed equally to this work and should be considered co-first authors.

References

[1] *Bruce, Fredrick. (1999). Factors contributing to recidivism among youth placed with the New York State Division for Youth.https://www.criminaljustice.ny.gov/crimnet/ojsa/dfy/dfy_research_report.pdf*

[2] *Dressel, J., & Farid, H. (2021). The Dangers of Risk Prediction in the Criminal Justice System. MIT Case Studies in Social and Ethical Responsibilities of Computing. https://doi.org/10.21428/2c646de5.f5896f9f*

[3] *Roberts, J. V. (1997). The role of criminal record in the sentencing process. Crime and Justice, 22, 303-362.*

DOI: 10.54254/2754-1169/17/20231066

[4] Latessa, E. J., Johnson, S. L., & Koetzle, D. (2020). What works (and doesn't) in reducing recidivism. Routledge.
[5] "Sigmoi." Sigmoi__Sumor, https://blog.csdn.net/su_mo/article/details/79281623.
[6] Oscar Torres-Reyna. (2007). Linear Regression using Stata https://www.princeton.edu/~otorres/Regression-101.pdf
[7] Weesie, J. (2001).Testing for Omitted Variables. https://www.stata.com/meeting/1nasug/weesie.pdf
[8] Maltz, M. D. (1984). Recidivism. Michael Maltz.
[9] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.
[10] AmiyaRanjanRout. (2022). Advantages and Disadvantages of Logistic Regression. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/