Reverse engineering secret algorithms: factors that influence parole decisions

Junjie Jiang '19
Mentor: Nathan Ryan, Vanessa Massaro
Bucknell University, Lewisburg, PA.

Abstract

We used web scraping and GET methods to acquire parole data from government websites. We then used Python and R to clean these data. We made sheets for data from different states or cities, including South Carolina and Virginia. Then, we used R to analyze the data to see the influence of different attributes on final parole decisions. Preliminary results according to logistic regression showed that race and sex do not appear to significantly influence final parole decisions. More careful and thorough analyses should be made for more accurate results.

Introduction

As computer programs are playing an increasing role in the correctional system, it is important to make sure that these programs are actually fair and sensitive attributes of criminals will not have any influence on the final decisions. The sensitive attributes include but are not limited to sex, race, age, zip code. These attributes are not relevant to the crime and should not be considered as reasons for sentences. Therefore, our research collects data and metadata of criminals along with their parole decisions and analyzes whether their parole decisions are made fairly according to mathematical and statistical analysis.

Background

In some states, the parole decisions are made by computer softwares that are provided by authorized third-party companies. With some information of criminals given to the software as input, the software will provide the decisions without transparency of how the decisions are made. Therefore, it is necessary for people without access to the black box algorithms to be assured of the equality within the system. The entire decision making process needs to ensure that no irrelevant attributes including but not limited to sex, race, age are of any influence.

Data Collection

There are several websites of different states that provides the public with parole information of criminals.
For South Carolina, since the state offers the data with the website, we used fetch queries to get the html file with different urls by changing the id included in the url. Then, we figured out the html elements that contain the information we need and stored them into a csv file.
For Virginia, the state provides the data in the form of pdf files online. Therefore, we used Tabula, a PDF file scraping software to acquire the data of the criminals and manually adjust some formatting errors.
After collecting the raw data, we used scripts to reorganize the data so as to facilitate the future analysis.

Data Analysis

Information from different states includes different sets of attributes. All of them include sex, race and age. Therefore, our data analysis basically focuses on the relationship between parole decisions and these attributes.
We used R to conduct our statistical analyses. First we removed the data points that are obvious outliers. Then we quantified all linguistic variables to numerical values. Finally we used logistic regression to statistically figure out if certain attributes are determinant for parole decisions.

```
> #logistic regression
> logis = glm(formula = Decision ~ Age + Race + Sex, data = newData, family = binomial)
> summary(logis)

Call:
glm(formula = Decision ~ Age + Race + Sex, family = binomial,
    data = newData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4363   0.3402   0.4951   0.5226   0.8597

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.385957   0.706212   0.547   0.5847
Age         0.010475   0.007559   1.386   0.1658
RaceWhite   0.836084   0.189406   4.414 1.01e-05 ***
SexMale     0.959785   0.568970   1.687   0.0916 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.6  on 1772  degrees of freedom
Residual deviance: 1150.5  on 1769  degrees of freedom
AIC: 1158.5

Number of Fisher Scoring iterations: 5
```

Analysis for Data from Virginia

Results

We did logistic regression analysis for two states seperately and did age, sex, race and three attributes together for each state.

- South Carolina
  - Age ~ Decision: $p < 0.001$
  - Sex ~ Decision: $p < 0.001$
  - Race ~ Decision: $p > 0.1$
  - Age + Sex + Race ~ Decision: Same as above
- Virginia
  - Age ~ Decision: $p < 0.05$
  - Sex ~ Decision: $p > 0.1$
  - Race ~ Decision: $p < 0.001$
  - Age + Sex + Race ~ Decision:
    - Age: $p > 0.1$
    - Sex: $p < 0.1$
    - Race: $p < 0.001$

Statistically, $p < 0.05$ represents that a certain attribute is significantly relevant. Therefore, for South Carolina, age and sex are significantly relevant and for Virginia, race is significantly relevant. Since the sample size of female is relatively small (124/1639 for SC and 23/1773 for VA), the influence of sex may be amplified.

Parole Decisions for October, 2017, with Reasons

DOC#	Name	Case Type	Decision Date	Decision	Certification Date	Age	Sex	Race	Not Grant Reasons Given / Conditions Violated
[redacted]	[redacted]	Violation Hearing	10/15/2017	Continue on Parole	10/17/2017	60	Male	Black	N/A
[redacted]	[redacted]	Violation Hearing	09/28/2017	Continue on Parole	10/02/2017	54	Male	Black	N/A
[redacted]	[redacted]	Violation Hearing	10/25/2017	Continue on Parole	10/26/2017	56	Male	White	N/A
[redacted]	[redacted]	Regular Parole	10/31/2017	Not Grant	10/31/2017	44	Male	White	Extensive criminal record
									Release at this time would diminish seriousness of crime
									The Board concludes that you should serve more of your sentence prior to release on parole.
									Your prior failure(s) and/or convictions while under community supervision indicate that you are unlikely to comply with conditions of release.
									Your record indicates a serious disregard for the property rights of others.
[redacted]	[redacted]	Regular Parole	10/17/2017	Not Grant	10/19/2017	60	Male	Black	No Interest in Parole
[redacted]	[redacted]	Regular Parole	10/07/2017	Not Grant	10/16/2017	49	Male	Black	Crimes committed - Sex Assault, Rape; Aggravated Sexual Battery; Malicious Wounding
									Release at this time would diminish seriousness of crime
									Serious nature and circumstances of your offense(s).
									The Board concludes that you should serve more of your sentence prior to release on parole.
									The Board considers you to be a risk to the community.

Sample Raw Data from Virginia

```
> logis = glm(formula = parole.decision ~ age + race + sex, data = newData, family = binomial)
> summary(logis)

Call:
glm(formula = parole.decision ~ age + race + sex, family = binomial,
    data = newData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2991   0.4115   0.5178   0.6066   1.1412

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.407074 308.987091   0.037    0.971
age           0.028518   0.005941   4.800 1.59e-06 ***
raceBLACK   -11.530754 308.986945  -0.037    0.970
raceOTHER   -12.297627 308.989124  -0.040    0.968
raceWHITE   -11.948689 308.986948  -0.039    0.969
sexMALE       0.981696   0.211495   4.642 3.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1420.4  on 1638  degrees of freedom
Residual deviance: 1357.7  on 1633  degrees of freedom
AIC: 1369.7

Number of Fisher Scoring iterations: 12
```

Analysis for Data from South Carolina

Conclusion

According to the preliminary results, it shows that in South Carolina and Virginia, attributes such as age, sex and race are correlated to the parole decision making. More specifically, people with less age, white color and female sex are more likely to receive positive parole decision.

Acknowledgements

Schotz Family Fund for Interdisciplinary Studies