Fatal Police Shootings in the United States

Brandeis University
International Business School

Team 5 H&A
Farheen Humayun & Angie Chen

Estimated annual rates of fatal police violence

12,209

People have been killed in the United States from 2000-2016 in police shootings

Increase of 38.24% in police shootings per 100,000 residents in recent years

Purpose of this report

Recent police shootings have lead to an outcry of systemic violence of state in the USA

Do police shootings disproportionately impact people of certain races or ethnicities, gender pointing to systemic biases in the police force?

Association of other factors such as mental illnesses or possession of a weapon? Is there a difference between behaviour across States of the US?

Data Source

- Compiled by Washington post.
- From the year 2000 to 2016.
- This database contains records of every fatal shooting in the United States by a Police officer in the line of duty.
- Washington Post claims that the dataset is updated regularly as facts emerge about every individual police shooting case.

1

Data Wrangling

A process to prepare data and improve data quality, making it ready for analysis.

Data Wrangling & Visualization

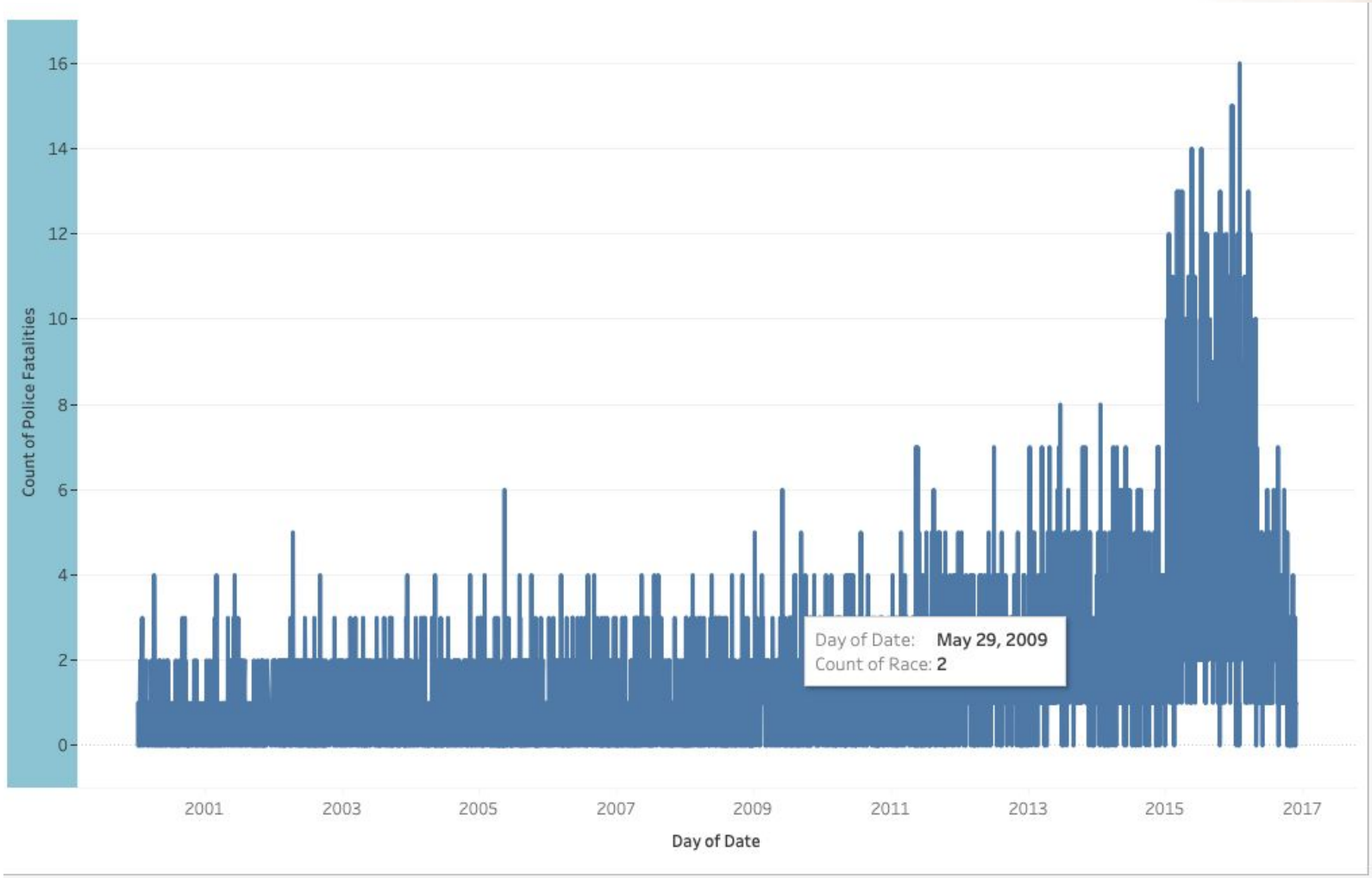
- We used tool Trifacta Wrangler
- **Imputation** - Imputing Missing Values using Mean Method
- **Normalization** - Delete columns
- **Transformation** - converting data to different format

  (Added Dummy variables , Extracted the year from the ‘Date’ column)

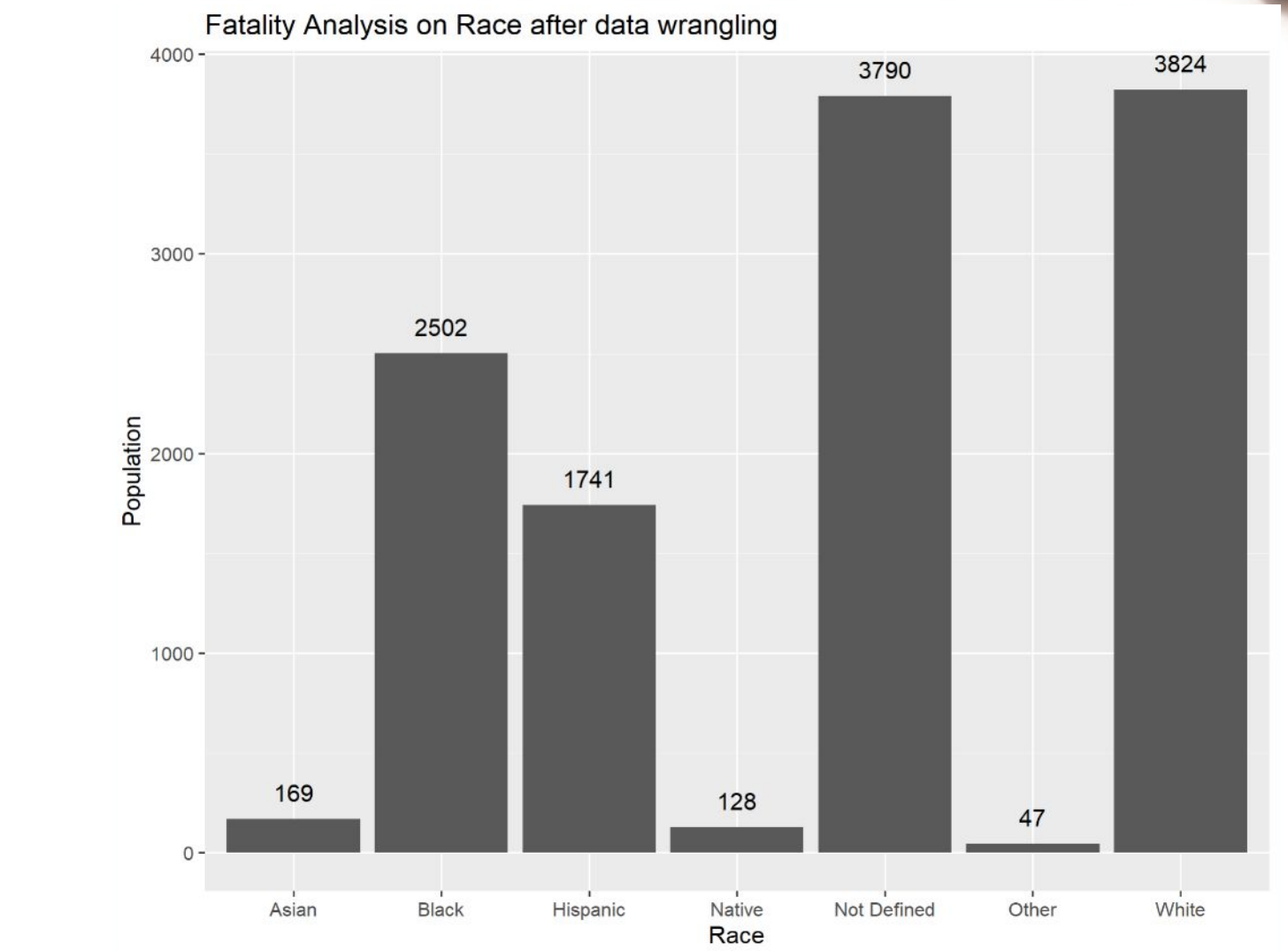
State	Manner_of_death	Armed	Mental_illness
51 Categories	4 Categories	59 Categories	2 Categories
CA	Shot		FALSE
CA	Shot		FALSE
CA	Shot	Gun	FALSE
CA	Shot	Gun	FALSE

Gender	Race	Date	City
2 Categories	6 Categories	2000 - 2016	3,316 Categories
Female	Asian	5/4/2000	Alameda
Male	Asian	6/2/2000	Fresno
Male	Asian	8/13/2000	Rosemead

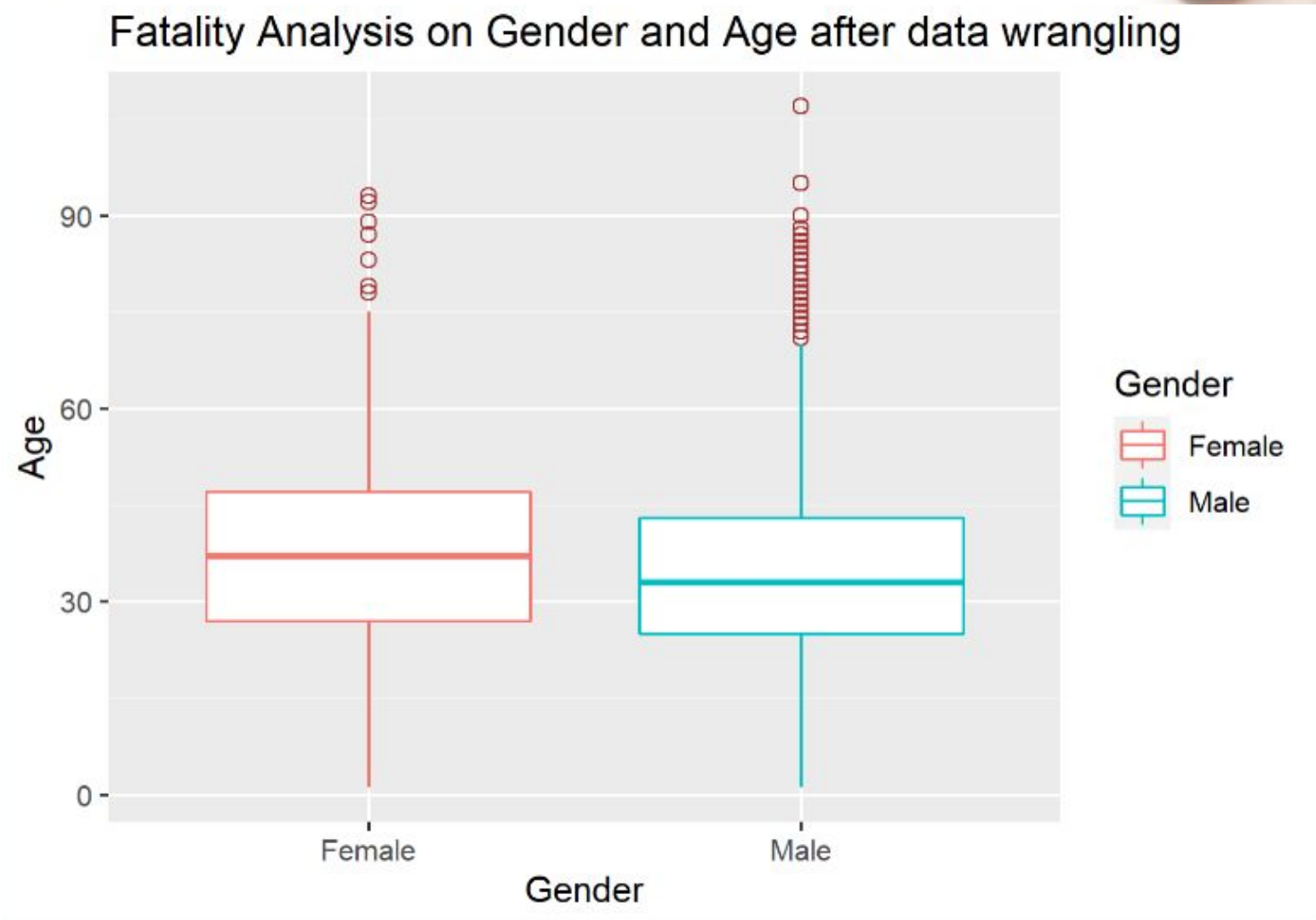
Police Fatalities - Year



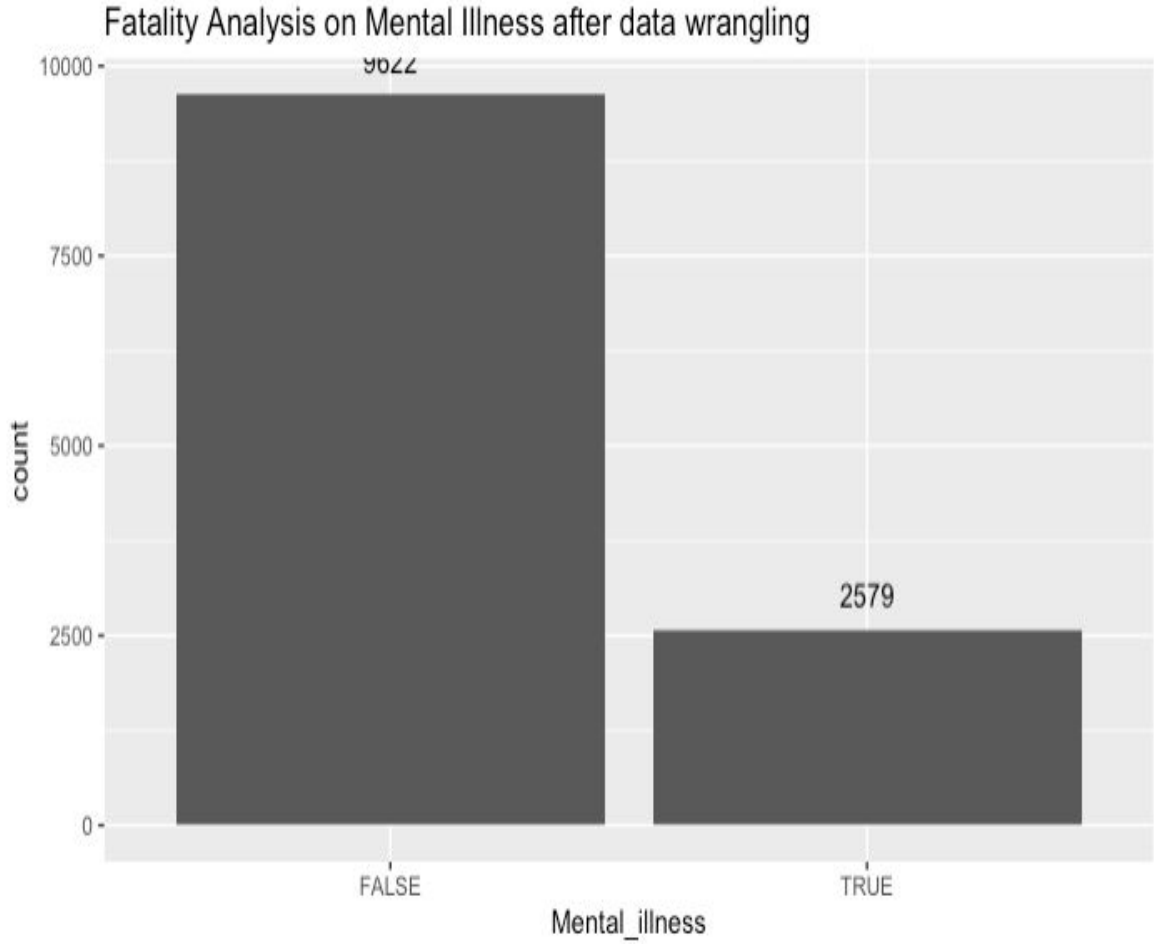
Police Fatalities - Year



Police Fatalities - Gender & Age

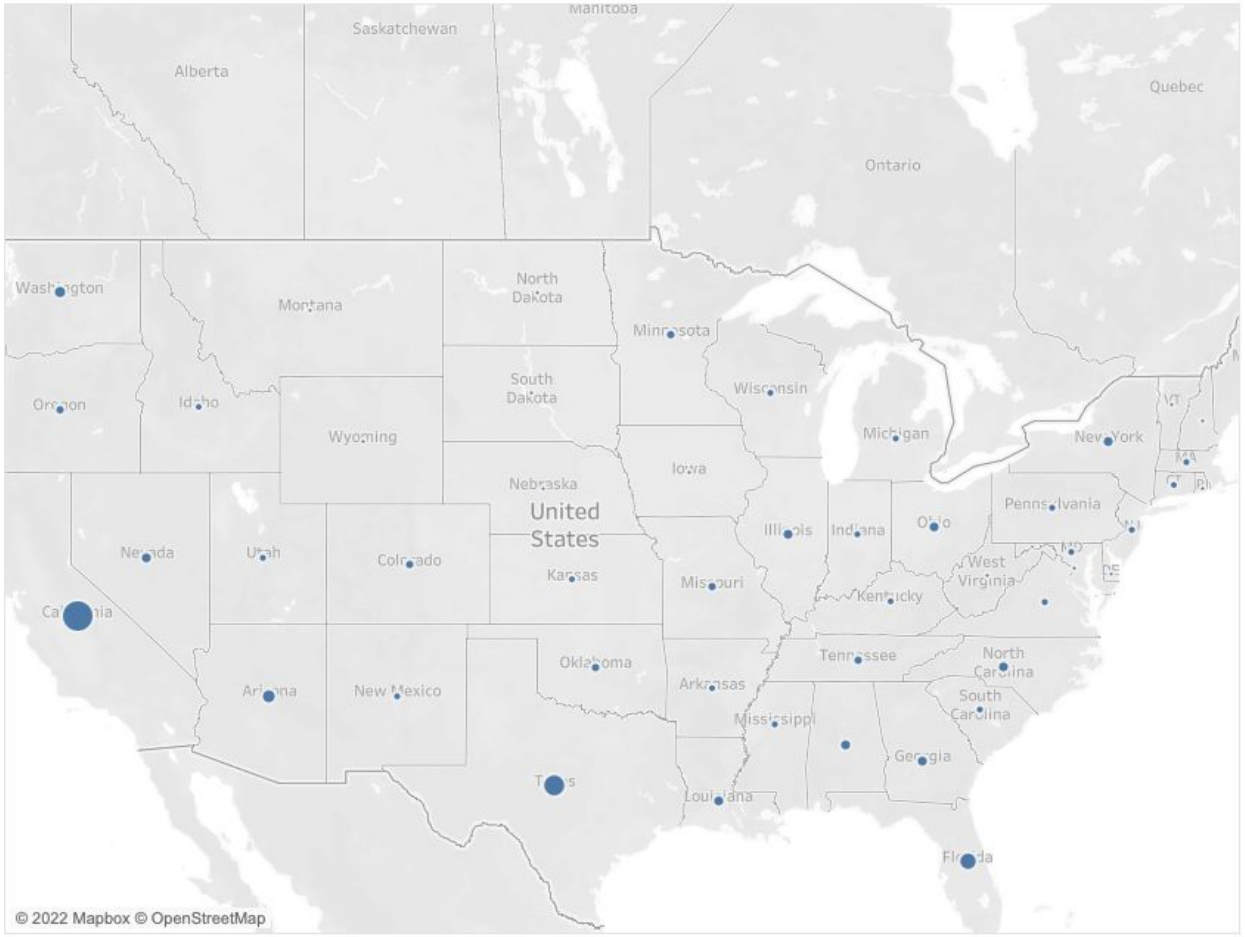


Police Fatalities - Mental Illness



Police Fatalities - State

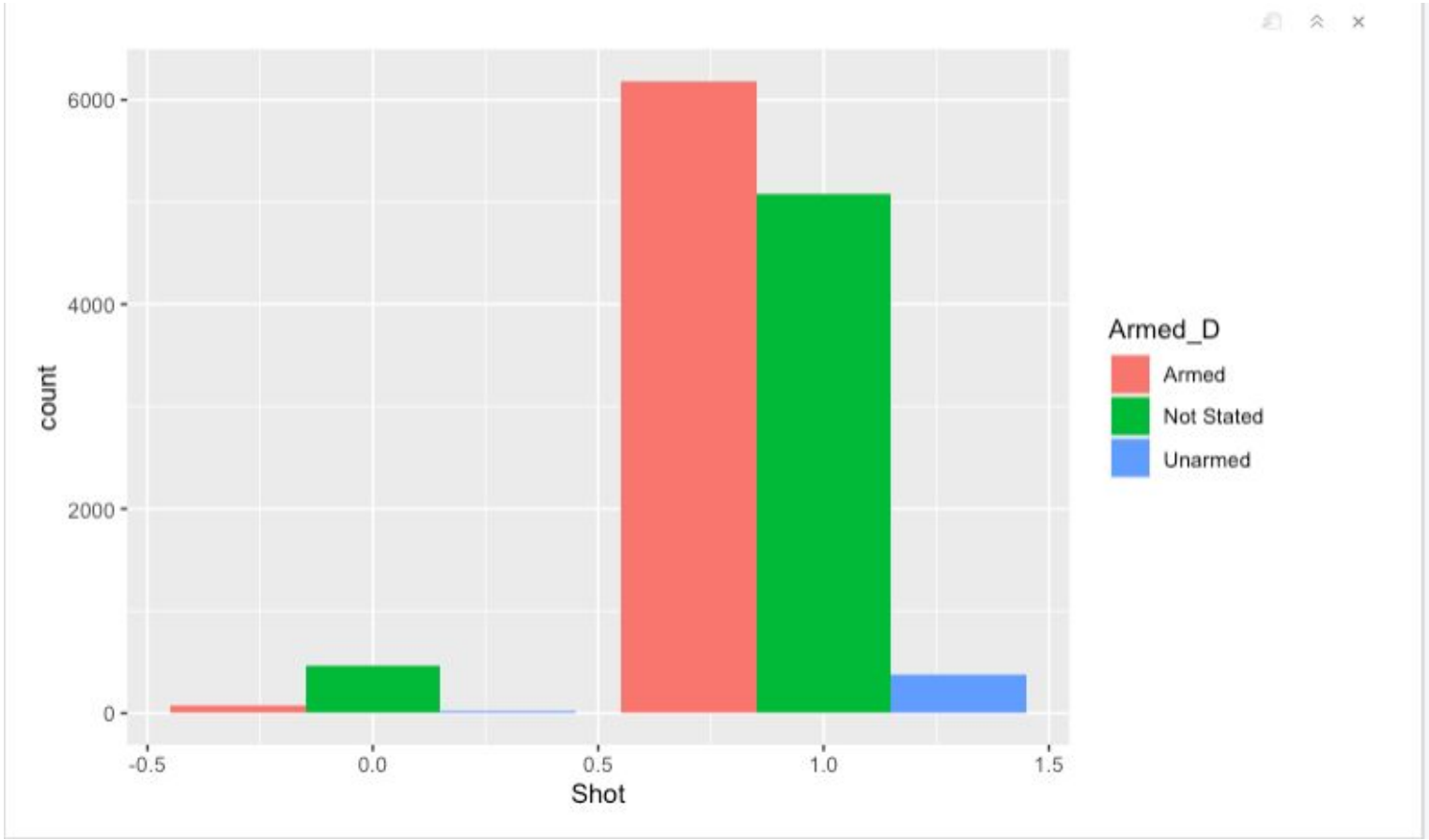
Count of Police Fatalities YoY



Count of Race
8
500
1,000
1,500
2,008

Map based on Longitude (generated) and Latitude (generated). Size shows count of Race. Details are shown for State.

Police Fatalities - Armed status



2

Regression Target

- Supervised learning
- Unsupervised learning
- Clustering
- Tree-Based Model

Target Variable for Regression Analysis

Multiple regression is used to analyze the relationship between a single dependent target and multiple explanatory predictors.

The dependent variable = (Fatal Rate)

The independent variables:

Age : Age at death.

Date : Violent Time (Month/Day/Year)

dummy_Mental_illness : (no illness =0, with illness =1)

Dummy_flee : (no flee = 0, try to flee =1)

PI_per_state : Average individual income per state

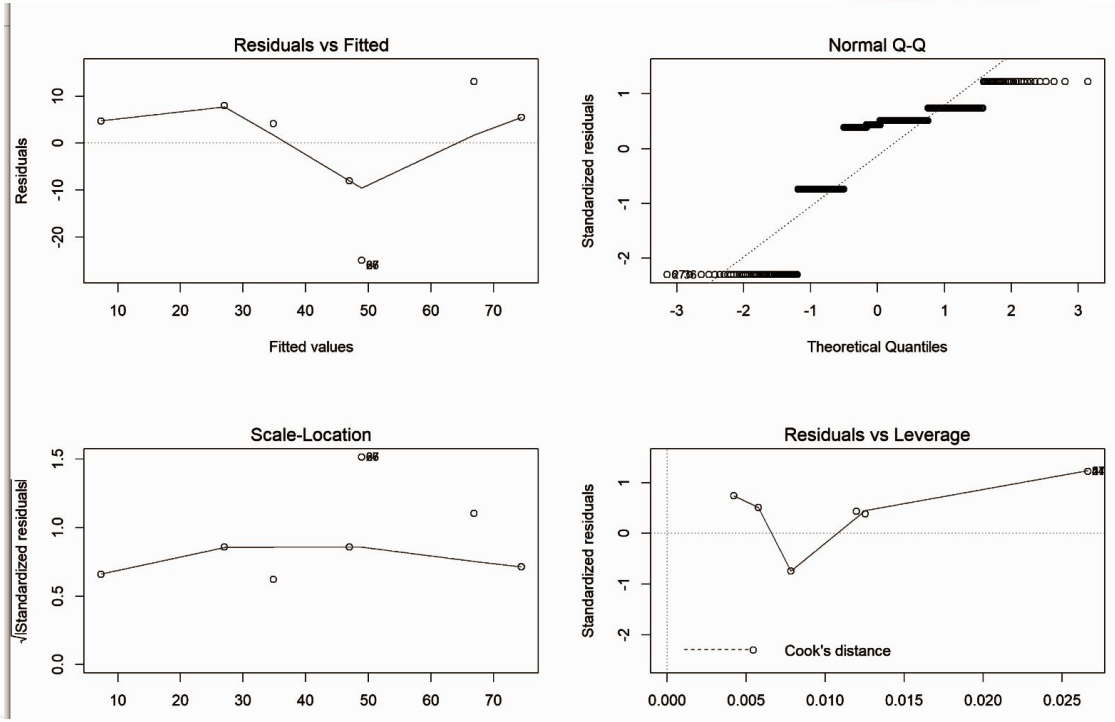
Crime_Case : Average Crime Case in every two years

Population : Population per each state

Homeless : Homeless per each state

Supervised Learning - Regression

- Fitting a Regression Initial Model
- **Remove 5 outliers**
- Diagnostic residual plots
- Forward / backward selection method
- polynomial term
- Accuracy Metrics
- **RMSE = 9.39**
- **R-squared = 0.844**

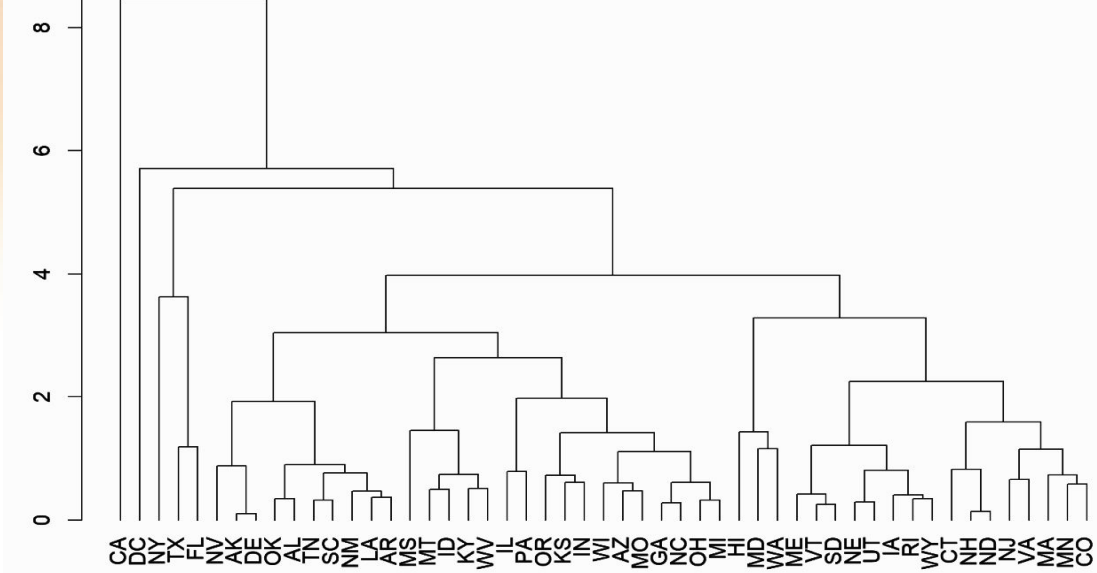



```
Multiple R-squared:  0.845,
F-statistic:  643 on 5 and
 Adjusted R-squared:  0.844
DF,  p-value: <0.000000000000
```

```
   ME  RMSE   MAE  MPE MAPE
0.19  9.39  7.38 -6.8 24.5
```

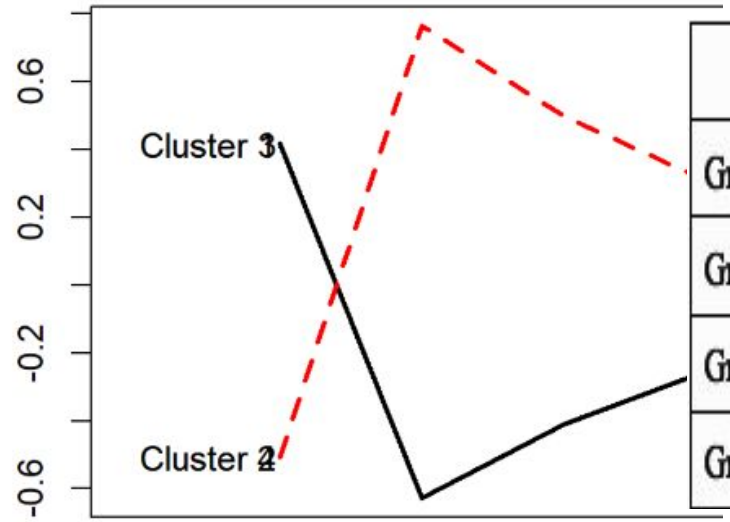
```
Coefficients:
                        Estimate   Std. Error t value             Pr(>|t|)
(Intercept)          491.492745104 35.491847240    13.8 < 0.0000000000000002 ***
PI_per_state          -0.000706814  0.000126214    -5.6          0.000000033 ***
Crime_Case            -0.646419655  0.045853842   -14.1 < 0.0000000000000002 ***
Population             0.000001403  0.000000359     3.9              0.00011 ***
Homeless               0.002274376  0.000087481    26.0 < 0.0000000000000002 ***
polynomial_CrimeCase   0.000218623  0.000014621    14.9 < 0.0000000000000002 ***
```

K-Means Clustering




```
> km$cluster
DC MD NJ MA CT MN WA CO NH HI ND VA AK DE IL NY RI WI WY CA NE IA VT NV PA SD KS UT OR TX AZ OH IN GA
 3  3  3  3  3  3  3  3  1  3  1  3  4  4  4  2  1  4  1  2  1  1  1  4  1  1  4  1  1  4  4  4  4  4
ME MI MO NC MT FL OK TN SC ID KY LA AL NM AR WV MS
 1  4  4  4  1  4  4  4  4  1  1  4  4  4  4  1  4
> cat(km$cluster)
3 3 3 3 3 3 3 3 1 3 1 3 4 4 4 2 1 4 1 2 1 1 1 4 1 1 4 1 1 4 4 4 4 4 1 4 4 4 1 4 4 4 4 1 1 4 4 4 4 1 4
> km$centers
```

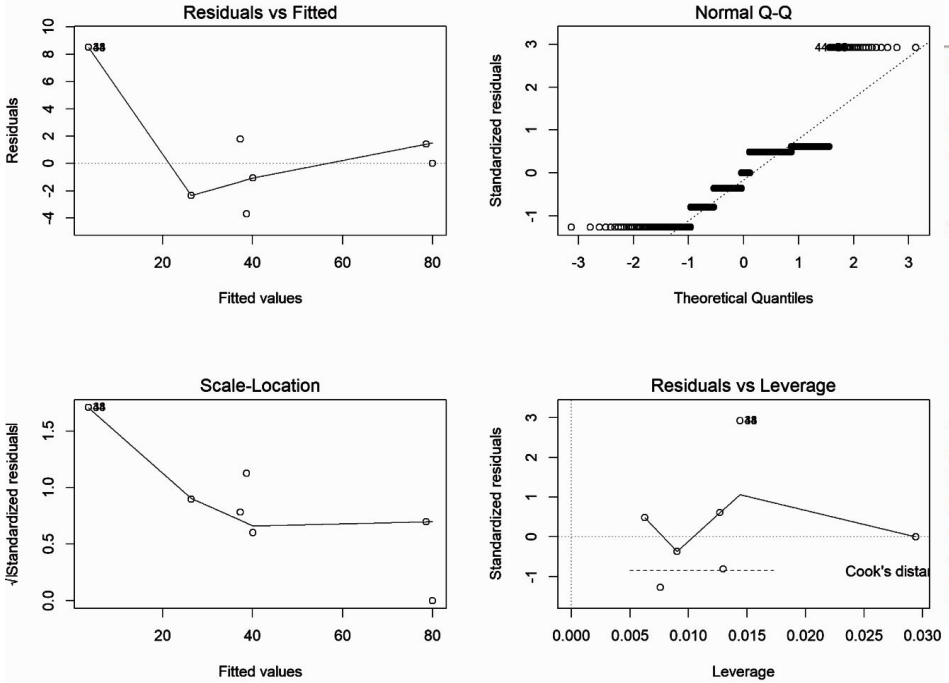
K-Means cluster labels



	PI per state	Crime case	Population	Homeless
Group1	25320-35807	950-1497	575251-12758729	548-15876
Group 2	31960-32320	1194-1701	19594330-38066920	92091-151278
Group 3	24509-49542	1125-1884	1392704-8874374	3033- 21577
Group 4	25758-34018	1584-2253	728300-26092033	921-28328

Unsupervised Learning in Regression

- Including polynomial terms and clusters
- Accuracy Metrics for the Final Regression Model
- **R-Sq =0.98**
- **RMSE = 2.97**
- Variance Inflation Factor




```
> summary(car.lm)

Call:
lm(formula = Fatal_police_shootings ~ . - Age - dummy_Mental_illness -
    dummy_flee - Crime_Case - Year, data = train.t)

Residuals:
   Min     1Q Median     3Q    Max
 -3.69  -2.33   0.00   1.42   8.50

Coefficients:
                        Estimate     Std. Error t value            Pr(>|t|)
(Intercept)          611.774791908   7.389093024    82.8 <0.0000000000000002 ***
PI_per_state          -0.010166712   0.000118787   -85.6 <0.0000000000000002 ***
Population             0.000007851   0.000000119    65.8 <0.0000000000000002 ***
Homeless              -0.000755771   0.000034355   -22.0 <0.0000000000000002 ***
polynomial_CrimeCase   0.000021288   0.000000183   116.0 <0.0000000000000002 ***
Cluster_Var         -139.205222509   1.626022957   -85.6 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

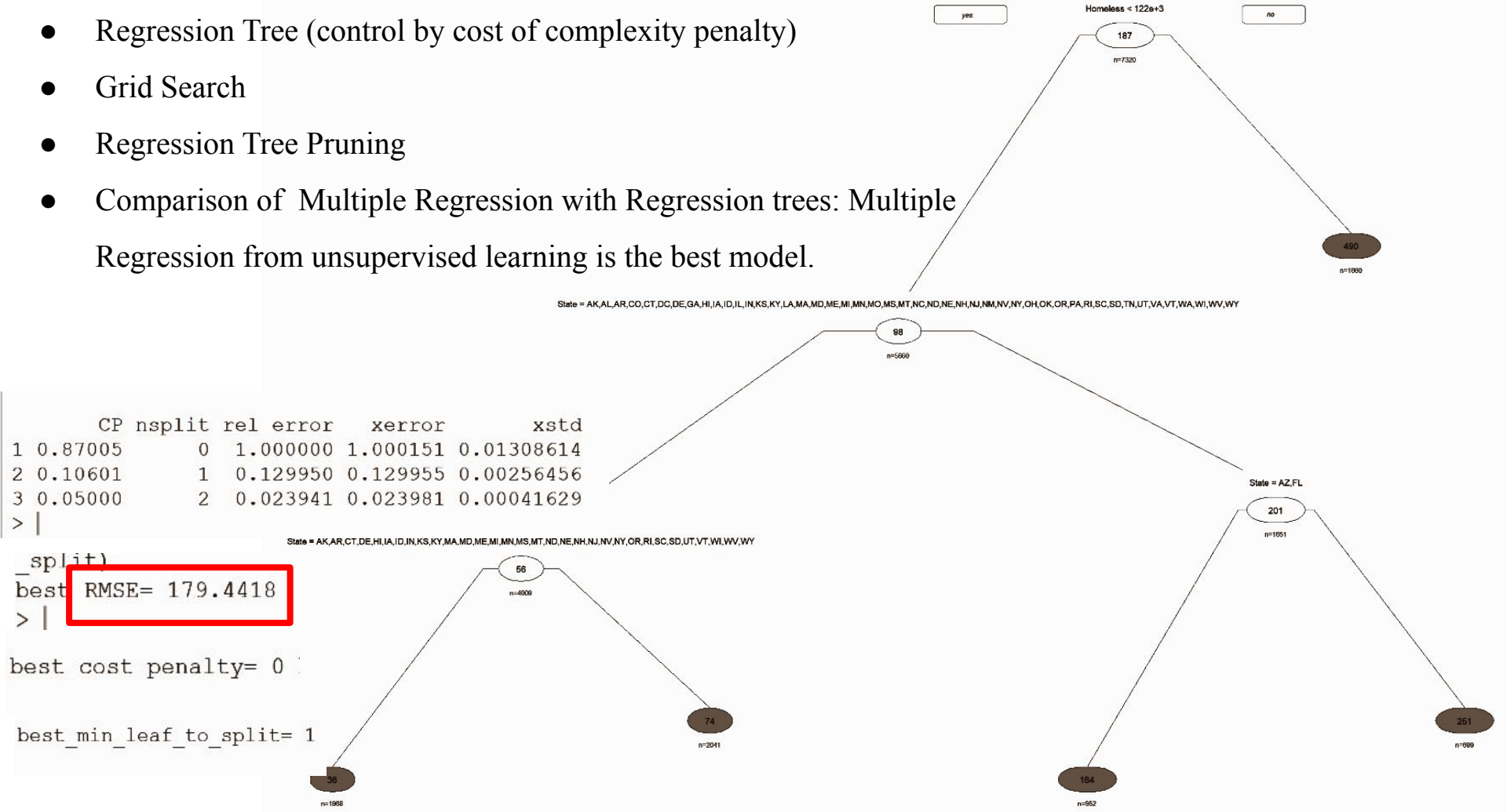
```
Multiple R-squared:  0.984,     Adjusted R-squared:  0.984
```

```
                      ME RMSE  MAE  MPE MAPE
set 0.0853 2.97 2.18 2.54  9.1
```

```
PI_per_state       Population         Homeless polynomial_CrimeCase
        9.96             1.50             3.75                 1.99
```

Tree-Based model - Regression Tree

- Regression Tree (control by cost of complexity penalty)
- Grid Search
- Regression Tree Pruning
- Comparison of Multiple Regression with Regression trees: Multiple Regression from unsupervised learning is the best model.




```
      CP nsplit rel error   xerror       xstd
1 0.87005      0  1.000000 1.000151 0.01308614
2 0.10601      1  0.129950 0.129955 0.00256456
3 0.05000      2  0.023941 0.023981 0.00041629
> |
```

```
_split)
best RMSE= 179.4418
> |
```

```
best cost penalty= 0
```

```
best_min_leaf_to_split= 1
```

3

Classification Target

- Supervised learning
- Tree-Based Model
- Ensemble Methods

Classification Model

Classification regression model is used to analyze the relationship between a single dependent target and multiple explanatory predictors.

The dependent variable = Were people killed by police through a gunshot or taser?

The independent variables:

1. Gender
2. Race
3. State
4. Do they suffer from a mental illness?
5. Did they try to flee from the police?
6. Year

Target Variable for Classification Analysis

Supervised Learning - Classification

- Race & whether they tried to flee or not were removed as they were found to be insignificant

- Gender, Mental Illness & Armed Status were found to be significant variables

```
Call:
glm(formula = Shot ~ Gender + Mental_illness + Armed_D, family = binomial(link = "logit"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4898   0.1192   0.2355   0.3551   0.6761

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)        6.08695    0.33308  18.275 < 0.0000000000000002 ***
GenderMale        -1.14287    0.31169  -3.667             0.000246 ***
Mental_illnessTRUE -1.37299    0.09108 -15.075 < 0.0000000000000002 ***
Armed_DNot Stated  -2.21172    0.12999 -17.014 < 0.0000000000000002 ***
Armed_DUnarmed     -1.54279    0.27213  -5.669         0.0000000143 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4551.1  on 12200  degrees of freedom
Residual deviance: 3926.5  on 12196  degrees of freedom
AIC: 3936.5

Number of Fisher Scoring iterations: 7

       (Intercept)         GenderMale Mental_illnessTRUE  Armed_DNot Stated      Armed_DUnarmed
       440.0769295          0.3189023          0.2533491          0.1095126          0.2137834
```

Supervised Learning - Classification

```
> confusionMatrix(prediction, traindata)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0    0    0
         1  342 6978

               Accuracy : 0.9533
                 95% CI : (0.9482, 0.958)
    No Information Rate : 0.9533
    P-Value [Acc > NIR] : 0.5144

                  Kappa : 0
```

```
Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.95328
             Prevalence : 0.04672
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 0
```

Accuracy: 95%

Very high specificity and a very sensitivity due to an imbalanced data set

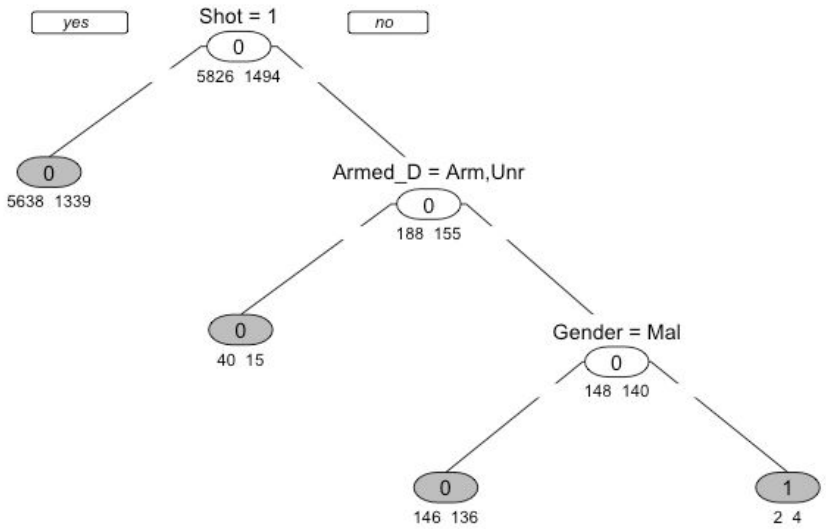
Tree-Based model - Classification Tree

The dependent variable = Are people suffering from a mental illness

The independent variables:

1. How was the person killed by the police
    a. Shot
    b. Taser etc
2. Gender
3. Armed status of the person in question
    a. Armed
    b. Unarmed
    c. Not stated in police reports

-



Cost penalty of the model: 0

Minimum leaf to split of the model: 8

F1 of the model: 18%

Ensemble Method - Boosting

The dependent variable = Are people suffering from a mental illness

The independent variables:

1. How was the person killed by the police
    a. Shot
    b. Taser etc
2. Gender
3. Armed status of the person in question
    a. Armed
    b. Unarmed
    c. Not stated in police reports

-

```
>
> # bagging (default parameters)
> for (mfinal0 in seq(1,2,3)) {
+   bag <- bagging(dummy_Mental_illness ~ ., data = train.df,mfinal = mfinal0)
+   pred <- predict(bag, valid.df, type = "class")
+   cm <- confusionMatrix(as.factor(pred$class), as.factor(valid.df$dummy_Mental_illness))
+
+   accuracy <- cm$overall[1]
+   P <- cm$byClass[1] # where did this come from?
+   R <- cm$byClass[2]
+   F1 <- (2 * P * R) / (P + R)
+
+   current_F1 <- (2*P*R) / (P+R)
+   if (current_F1 > best_F1) {
+     best_F1 <- current_F1
+     bestmfinal <- mfinal0
+   }}
>
> cat("bagging trees accuracy=", accuracy, " and ", " F1=", F1)
bagging trees accuracy= 0.7744315  and   F1= 0.1434337> print(current_F1)
Sensitivity
  0.1434337
>
```

Accuracy: 77%

F1 of the model: 14%

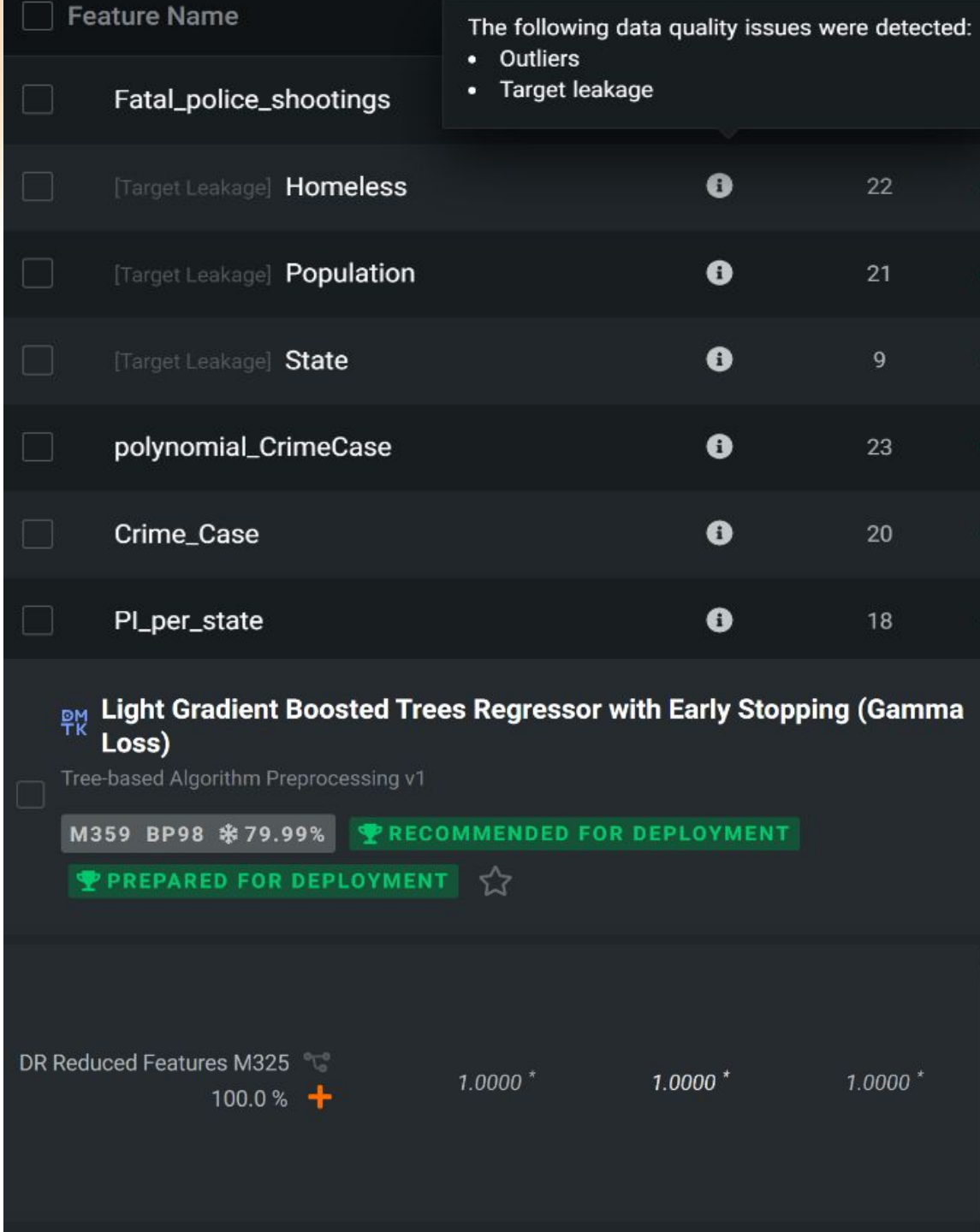
4 DataRobot Models (Regression)

Target Leakage

Achieving performance that seems a little too good to be true.

- **Temporal Cutoff.** Remove all data just prior to the event of interest, focusing on the time you learned about a fact or observation rather than the time the observation occurred.

- **Remove Leaky Variables.** Evaluate simple rule based models line OneR using variables like account numbers and IDs and the like to see if these variables are leaky, and if so, remove them. If you suspect a variable is leaky, consider removing it.



#1 Regression Model in DataRobot

Ridge Regressor

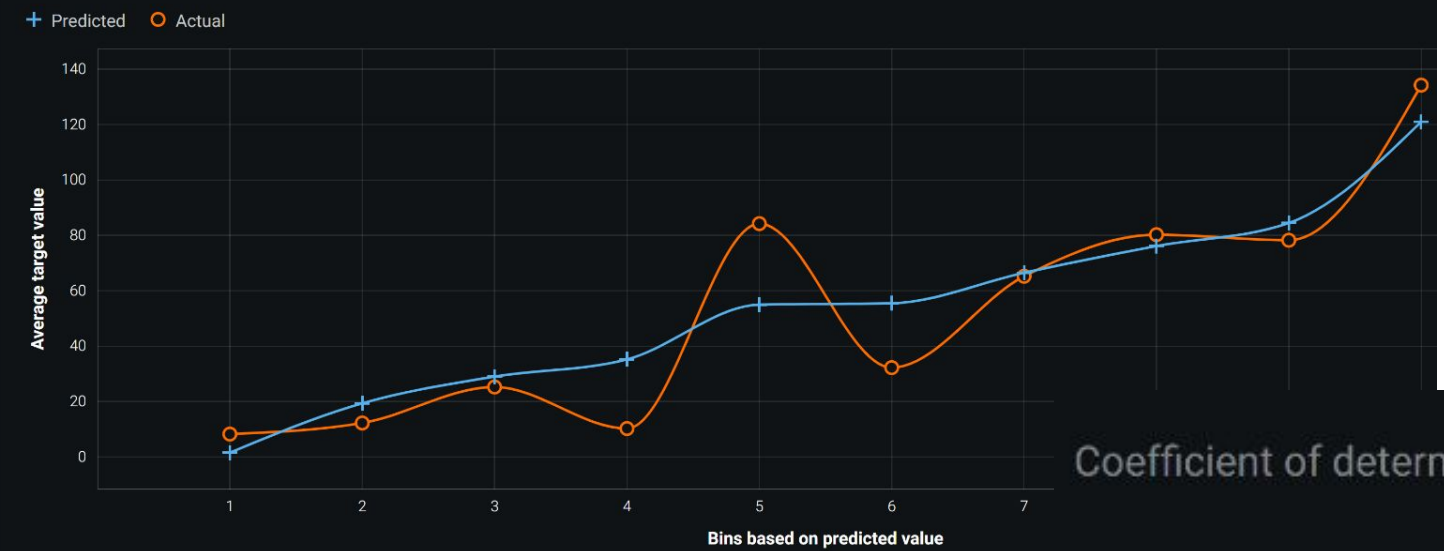
Missing Values Imputed | Standardize | Ridge Regressor

M57 BP3 REF β_i

Informative Features

62.75 %

	R-Squared	RMSE
	0.5979	19.1220
	0.4598	35.1083
	0.8927	15.8787



Target = Fatal police shooting rate

Coefficient of determination (r^2): 0.8927

#2 Regression Model in DataRobot

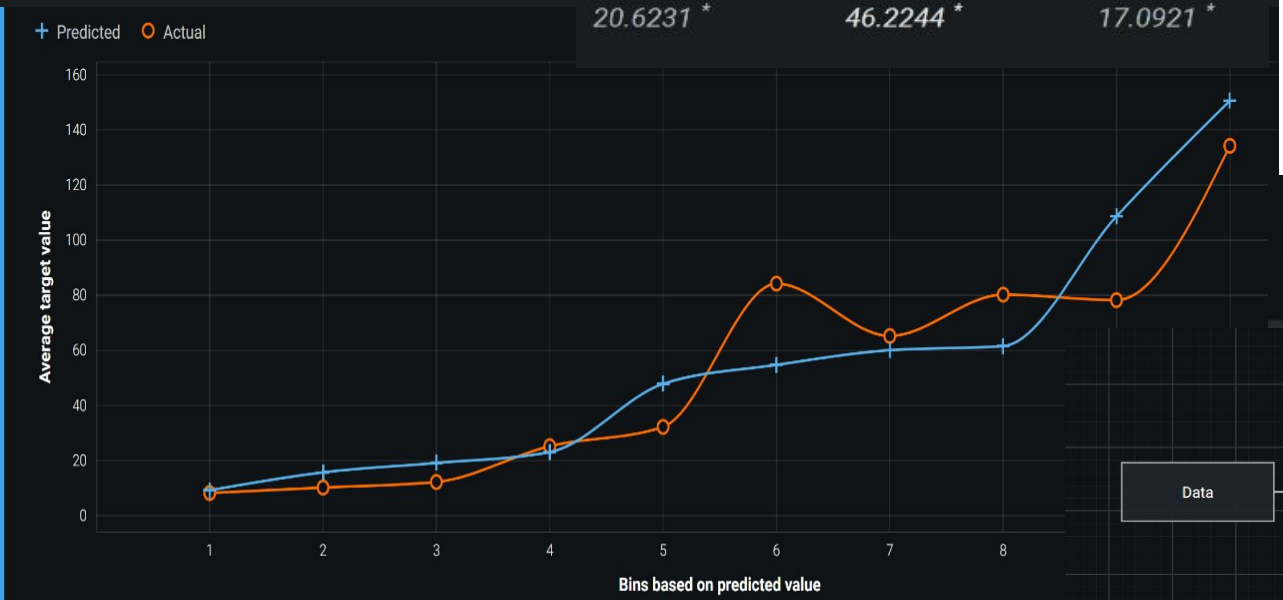
Elastic-Net Regressor (mixing alpha=0.5 / Gamma Deviance)

Regularized Linear Model Preprocessing v2

M42 BP69 78.43% RECOMMENDED FOR DEPLOYMENT

PREPARED FOR DEPLOYMENT

	R-Squared	RMSE
Informative Features	0.5323 *	20.6231 *
100.0 %	0.1439 *	46.2244 *
	0.8757 *	17.0921 *



Target = Fatal police shooting rate

Coefficient of determination (r^2): 0.8757

Own Models vs. DataRobot

(Regression Analysis)

Unsupervised Learning - Regression

- RMSE 2.97
- R-squared 0.984

Supervised Learning - Regression

- RMSE 9.4
- R-squared 0.84

Regression Tree

- RMSE 179

Ridge Regressor

- RMSE 15
- R-squared 0.89

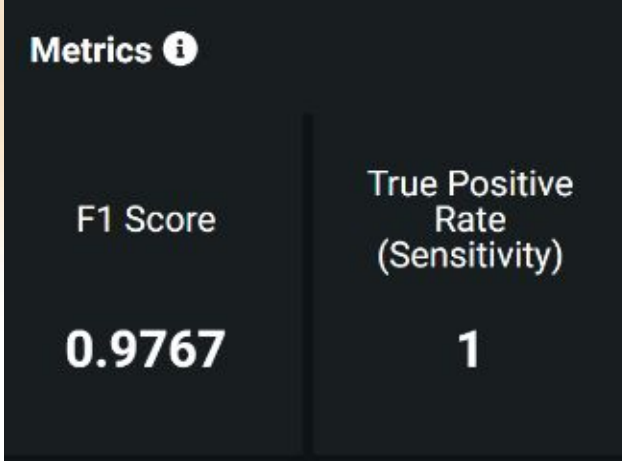
Elastic-Net Regressor

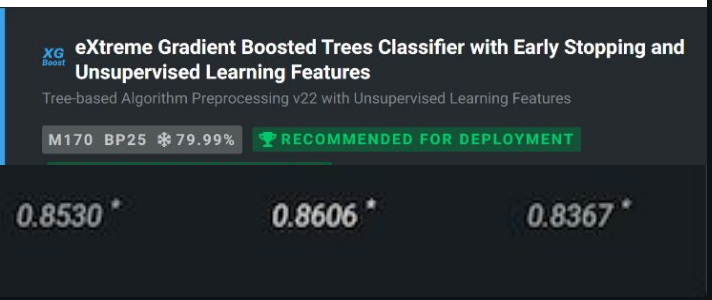
- RMSE 17
- R-squared 0.87

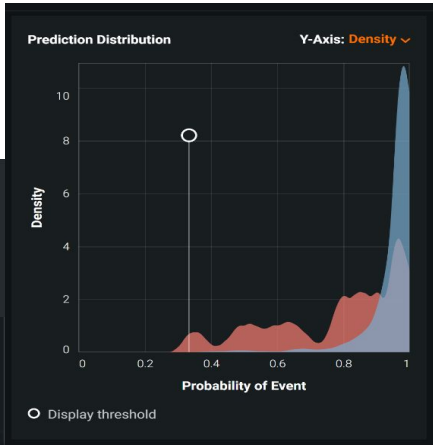
5 DataRobot Models (Classification)

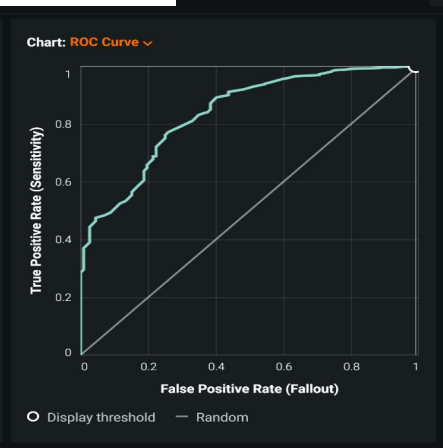
Data Robot - Classification Model 1

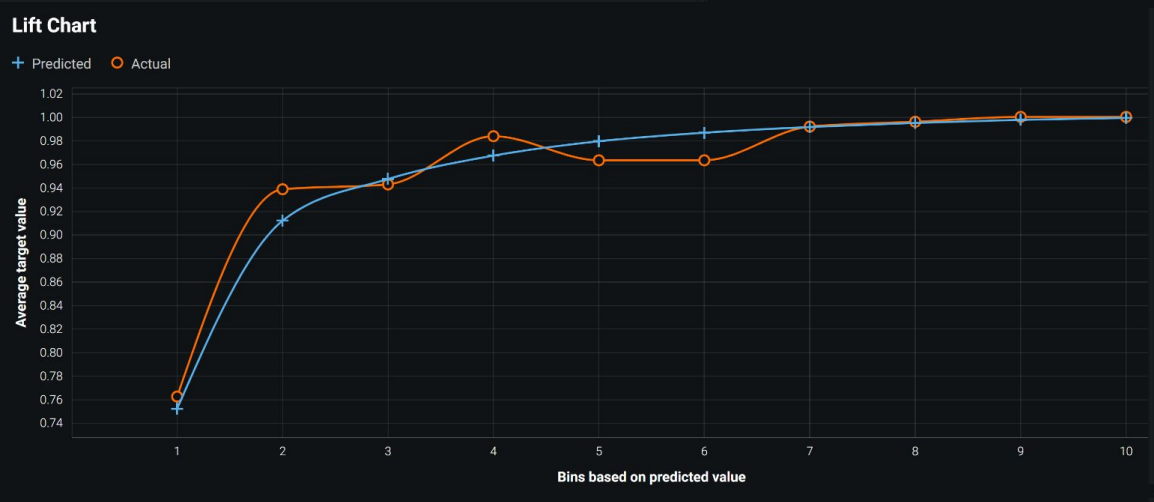
Target Variable - Were people shot or killed?





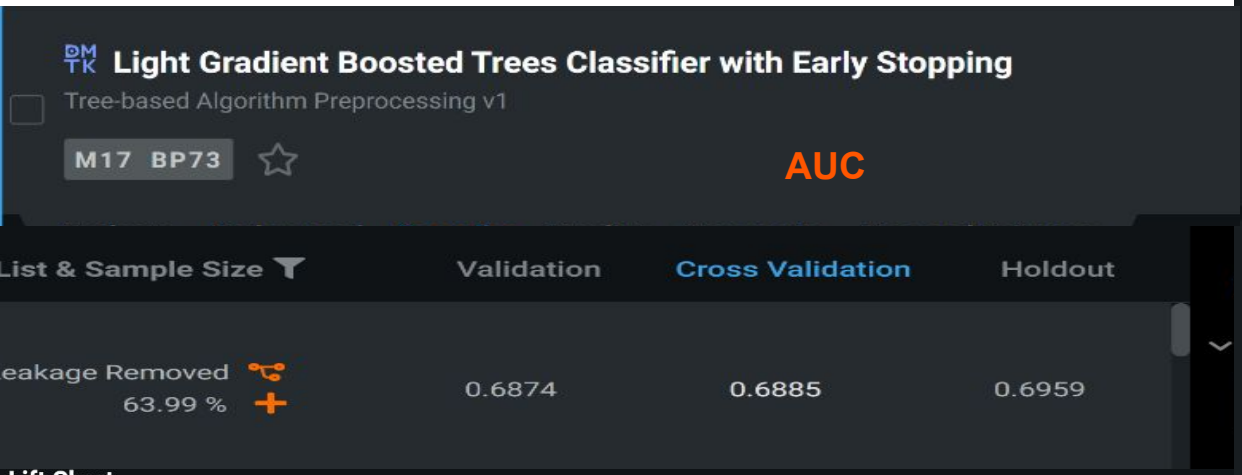


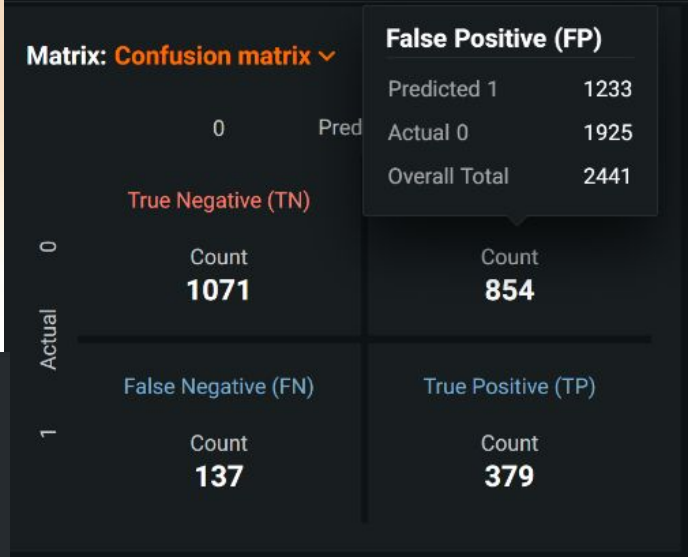


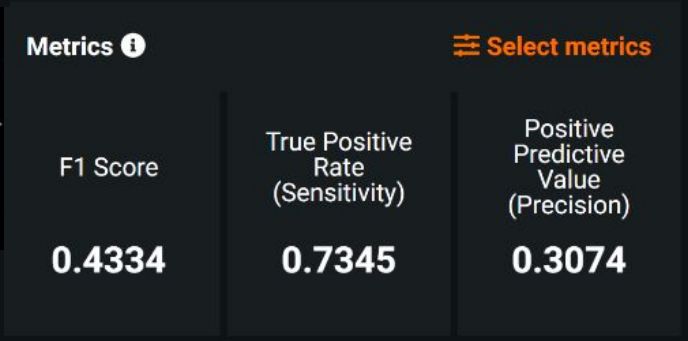


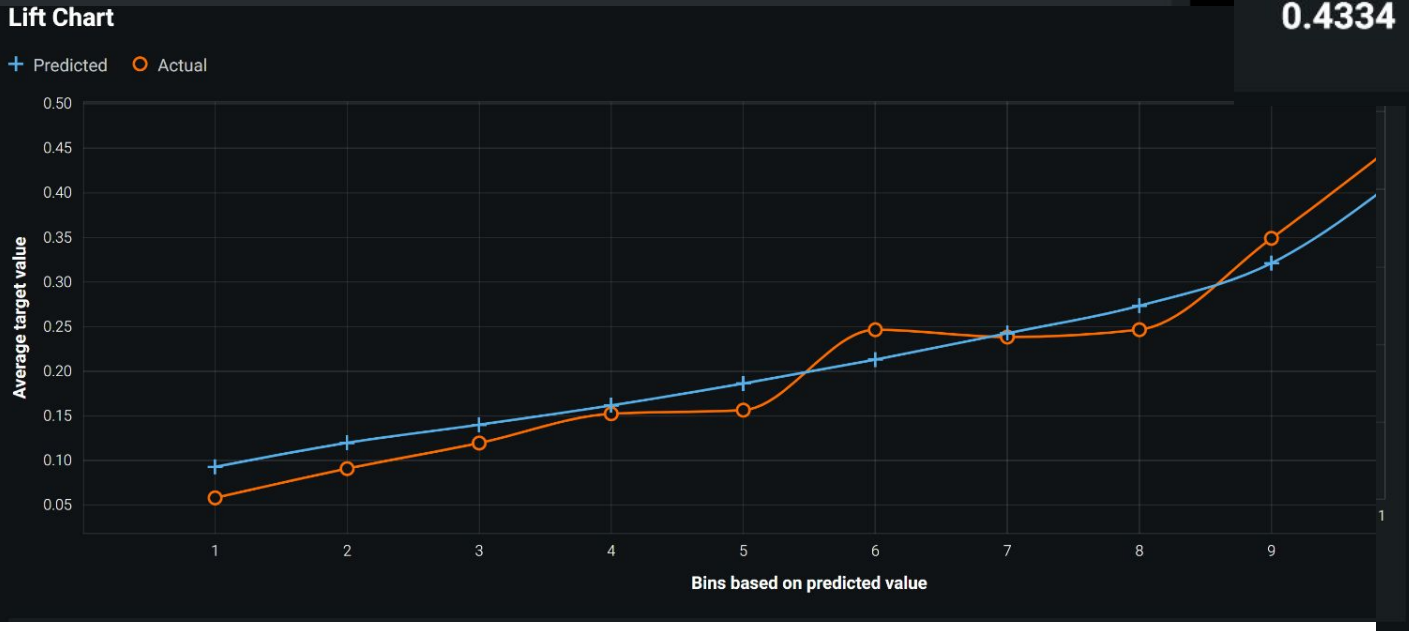
Data Robot - Classification model 2

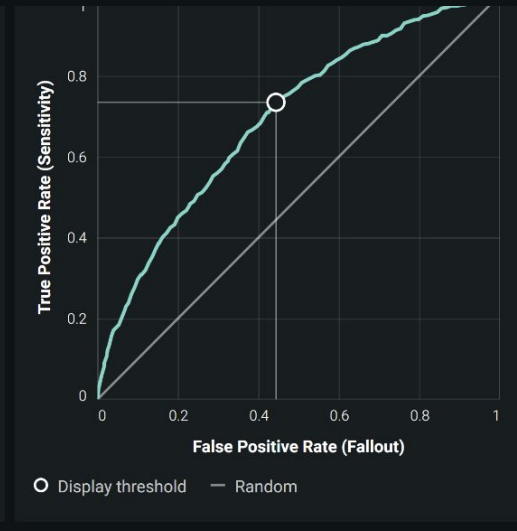
Target Variable - Do people suffer from a mental illness?











Conclusions and Insights

Imbalanced data set wrt. our target variables. Under sampling or over sampling could have been used to balance the data set and to obtain better sensitivity and specificity

Our regression analysis show **mental illness** and the **armed status** have strong association with whether will be shot or tasered

Personal income per state, Crime Case, Population and number of Homeless have directly impact on predicting the rate of fatal police shooting. Datarobot determined those variables as outliers and target leakage. But models generates from human have higher accuracy in preventing those issues.

Thank You!

Any questions?

Citation

Published by Statista Research Department, & 1, F. (n.d.). *People shot to death by U.S. police, by race from 2007 to 2022*. Statista. Retrieved from https://www.statista.com/statistics/585152/people-shot-to-death-by-us-police-by-race/

Mapping Police Violence. (n.d.). Retrieved February 3, 2022, from https://mappingpoliceviolence.org/

Guardian News and Media. (n.d.). *The counted: Tracking people killed by police in the United States | US news*. The Guardian. Retrieved February 3, 2022, from https://www.theguardian.com/us-news/series/counted-us-police-killings

Larxene. (n.d.). Fatal Police Shootings in the US (2015-2020). https://www.kaggle.com/andrewmvd/police-deadly-force-usage-us

Fatal police violence by race and state in the USA, 1980–2019: a network meta-regression. The Lancet Journal. (n.d.). Retrieved October 2, 2021, from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)01609-3/fulltext

Hemenway, D., Azrael, D., Conner, A., & Miller, M. (2019, February). *Variation in rates of fatal police shootings across US states: The role of Firearm availability*. Journal of urban health : bulletin of the New York Academy of Medicine. Retrieved March 4, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6391295/

U.S. crime index state rank. USA.com. (n.d.). Retrieved March 4, 2022, from http://www.usa.com/rank/us--crime-index--state-rank.htm

Homeless population by state 2022. (n.d.). Retrieved March 4, 2022, from https://worldpopulationreview.com/state-rankings/homeless-population-by-state