Project Report

Rishabh Bhatia

OVERVIEW

Racism and police violence are long-standing problems in the United States. The recent killing of George Floyd at the hands of the Minneapolis PD is the latest example. It has received the entire world's attention and brought the issue front and center once again. This project examines to what degree the policies of a police department relate to its officers' use of deadly force. The dataset that I've chosen is publicly available on Kaggle.com and is compiled by JPMiller. This dataset contains data from the Police Use of Force project, Mapping Police Violence, and the Washington Post.

INTRODUCTION

CAVEAT

The analysis presented in this notebook is my interpretation of the datasets provided. Much of the articles that have come in the media in recent times shed light on a singular view of this very broad, multi-faceted topic. I have attempted to not be biased in this report and hence, my views and points might attempt to represent multiple sides of the broad issue. This is predominantly because data itself is open to interpretations. The topic at hand is largely engrossed in a human-centred outlook and hence it becomes very difficult to quantify it with absolute precision. Hence, I do apologize for any mistakes that could have creeped into the analysis due to my lack of first-person exposure with the domain. In case of a discrepancy that has occured due to my work, please do let me know of the same.

DATASET DESCRIPTION

The dataset contains 2 datasets: 1. Washington Post Shooting Data, 2. Deaths and Arrests Dataset

The above graph shows the deaths of Black people caused by police between 2013-2019. It shows that Texas is the state where most Black people deaths occur, followed by California in second most largest number of Black people deaths.

The above graph shows the deaths of Hispanic people caused by police between 2013-2019. It shows that California is the state where most Hispanic people deaths occur, followed by Texas in second most largest number of Hispanic people deaths.

The above graph shows the deaths of Native American people caused by police between 2013-2019. It shows that Arizona is the state where most Native American deaths occur, followed by Oklahoma and Washington in second most largest number of Native American people deaths.

The above graph shows the deaths of Asian people caused by police between 2013-2019. It shows that California is the state where most Asian people deaths occur, followed by Texas in second most largest number of Asian people deaths.

The above graph shows the deaths of Pacific Islanders caused by police between 2013-2019. It shows that Hawaii is the state where most Pacific Islanders deaths occur, followed by California in second most largest number of Pacific Islanders deaths.

The above graph shows the deaths of White people caused by police between 2013-2019. It shows that California is the state where most White people deaths occur, followed by Arizona in second most largest number of Pacific Islanders deaths.

The above graph shows the percentage Black people deaths caused by police between 2013-2019 in the state mentioned. It shows that Louisiana is the state where the highest percentage of black deaths happen per state, followed by Maryland in second most largest percentage Black people deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the percentage of Hispanic people deaths caused by police between 2013-2019 in the state mentioned. It shows that New Mexico is the state where the highest percentage of Hispanic deaths happen per state, followed by California in second most largest percentage of Hispanic people deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the percentage of Native American deaths caused by police between 2013-2019 in the state mentioned. It shows that Alaska and Minnesota is the state where the highest percentage of Native American deaths happen per state, followed by Washington in second most largest percentage of Native American deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the percentage of Asian people deaths caused by police between 2013-2019 in the state mentioned. It shows that North Carolina is the state where the highest percentage of Asian people

deaths happen per state, followed by New York in second most largest percentage of Asian people deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the percentage of Pacific Islanders deaths caused by police between 2013-2019 in the state mentioned. It shows that Hawaii is the state where the highest percentage of Pacific Islanders deaths happen per state, followed by North Carolina in second most largest percentage of Pacific Islanders deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the percentage of White people deaths caused by police between 2013-2019 in the state mentioned. It shows that Kansas is the state where the highest percentage of White people deaths happen per state, followed by Oregon in second most largest percentage of White people deaths caused by police between 2013-2019 in the state mentioned.

The above graph shows the total deaths that occur per month overall as a total in the five years between 2015-2020. The dip in months from July to December are caused due to no data for these months in 2020. The total for months from July to December could be higher by roughly 50-80 deaths.

The above graph shows the total deaths that occurred per year in the five years between 2015-2020. The dip in 2020 total deaths is partly because of data only till July and because of COVID-19 complications.

The above graph shows the total deaths that occurred per weekday overall in the five years between 2015-2020.

The above graph shows the total deaths that occurred per race per year in the five years between 2015-2020. The dip in 2020 is caused due to incomplete data and COVID-19 complications and social distancing. The orange line without a race class is those of immigrants.

The above graph shows the total deaths that occurred per State in the five years between 2015-2020. We can see that California has the most number of victims. Texas and Florida are next to California. These three are densely populated states in US.

The above graph shows the top 10 cities with the highest total deaths that occurred in the five years between 2015-2020. Los Angeles, Phoenix and Houston have recorded most victims.

Summary of preliminary conclusions According to my exploratory data analysis, the following are my preliminary conclusions:

1. As we observed from previous graph, January, February and March have recorded most cases.
2. This graph shows that the rate of killing per year has never looked down except 2016.
3. It has an average of 983 per year. Almost 1000 people are killed every year.
4. We are in mid of 2020 and we have already reached 432 death counts as of now.
5. Most people are killed around mid of weekdays
6. Weekends proportion to weekdays are relatively smaller.
7. We could see more of white race and black people are affected. Immigrants are facing more deaths increase per year. Mid way though 2020 they've already risen past last year's numbers.
8. California has recorded most number of victims.
9. Texas and Florida are next to California. These three are densely populated states in US.
10. Average number of victims per state is 106
11. Densely populated state are the most victim faced state.
12. Los Angeles, Phoenix and Houston have recorded most victims. 13. White and Black peoples are the most affected ones. 14. Hispanic victims are lying close to series of black people. 15. There is a drop in early 2019 but it lifted back in December 2019.

Model 1 SVM Models

Linear Kernel

The Support Vector Machine is a supervised learning algorithm mostly used for classification but it can be used also for regression. The main idea is that based on the labeled data (training data) the algorithm tries to find the optimal hyperplane which can be used to classify new data points. In two dimensions the hyperplane is a simple line.

```
##
## Call:
## svm(formula = threat_level ~ ., data = shooting.train, kernel = "linear",
##     cost = 0.01)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.01
##
## Number of Support Vectors:  3380
##
##  ( 1902 1298 180 )
##
##
## Number of Classes:  3
##
## Levels:
##  attack other undetermined
```

In the above model total number of support vectors used are 3380. Out of which 1902 are of class "Attack", 1298 of class "Other" and "180" of class "Undetermined".

Next, we finding the training classification error rate and the test classification error for Linear Kernel

```
## [1] "Training classification error rate is 0.398379473328832"
```

```
## [1] "Test classification error rate is 0.385933273219116"
```

From out Linear Kernel SVM Model, we can see that Training Classification Error Rate is 39.83% and Test Classification Error is 38.59%.

The Second SVM model we run is Polynomial.

```
##
## Call:
## svm(formula = threat_level ~ ., data = shooting.train, kernel = "poly",
##     cost = 0.01)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  polynomial
```

```
##        cost:  0.01
##      degree:  3
##      coef.0:  0
##
## Number of Support Vectors:  2807
##
##  ( 1329 1298 180 )
##
##
## Number of Classes:  3
##
## Levels:
##  attack other undetermined
```

In the above model total number of support vectors used are 2807. Out of which 1329 are of class "Attack", 1298 of class "Other" and "180" of class "Undetermined".

Next, we finding the training classification error rate and the test classification error for Polynomial Kernel

```
## [1] "Training classification error rate is 0.353364843574162"
```

```
## [1] "Test classification error rate is 0.352569882777277"
```

From out Polynomial Kernel SVM Model, we can see that Training Classification Error Rate is 35.33% and Test Classification Error is 35.25%. We can see that Polynomial Model is better than Linear Kernel Model.

The Third SVM model we run is Radial.

```
##
## Call:
## svm(formula = threat_level ~ ., data = shooting.train, kernel = "radial",
##     cost = 0.01)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  0.01
##
## Number of Support Vectors:  2801
##
##  ( 1323 1298 180 )
##
##
## Number of Classes:  3
##
## Levels:
##  attack other undetermined
```

In the above model total number of support vectors used are 2801. Out of which 1323 are of class "Attack", 1298 of class "Other" and "180" of class "Undetermined".

Next, we finding the training classification error rate and the test classification error for Radial Kernel

```
## [1] "Training classification error rate is 0.353364843574162"
```

```
## [1] "Test classification error rate is 0.352569882777277"
```

From our Radial Kernel SVM Model, we can see that Training Classification Error Rate is 35.33% and Test Classification Error is 35.25%. We can see that Radial Model is better than Linear Kernel Model.

```
##
## Call:
## best.tune(method = svm, train.x = threat_level ~ ., data = shooting.train,
##     ranges = list(cost = c(seq(0.01, 0.1, 1))), kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  0.01
##
## Number of Support Vectors:  2801
##
##  ( 1323 1298 180 )
##
##
## Number of Classes:  3
##
## Levels:
##  attack other undetermined
```

In the above model total number of support vectors used are 2801. Out of which 1323 are of class "Attack", 1298 of class "Other" and "180" of class "Undetermined".

This is our final Tuned model.

Model 2 Naives Bayes Classifier

The Second Model I ran is the Naives Bayes Classifiers, Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

```
P(A|B) = \frac{P(B|A) P(A)}{P(B)}

 where A and B are events
```

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence. P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).

P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##       attack        other undetermined
##   0.64663516   0.30880036   0.04456448
##
## Conditional probabilities:
##               manner_of_death
## Y                    shot shot and Tasered
##   attack       0.96310477       0.03689523
##   other        0.92274052       0.07725948
##   undetermined 0.96464646       0.03535354
##
##               signs_of_mental_illness
## Y                  False      True
##   attack       0.7894187 0.2105813
##   other        0.7470845 0.2529155
##   undetermined 0.8434343 0.1565657
##
##               gender
## Y                                  F           M
##   attack       0.000000000 0.040375914 0.959624086
##   other        0.000728863 0.053935860 0.945335277
##   undetermined 0.000000000 0.030303030 0.969696970
##
##               race
## Y                                  A           B           H           N
##   attack       0.110685694 0.013574661 0.246084233 0.143056039 0.012878524
##   other        0.088921283 0.023323615 0.218658892 0.189504373 0.015306122
##   undetermined 0.156565657 0.005050505 0.222222222 0.207070707 0.015151515
##               race
## Y                         O           W
##   attack       0.009049774 0.464671076
##   other        0.009475219 0.454810496
##   undetermined 0.000000000 0.393939394
```

In this project, I used the Naïve Bayes Classifier to find the conditional probabilities between variables. That is, condition probabilities between Threat Level and multiple other variables such as Manner of Death, Signs of Mental Illness, Gender and Race.

Results from Naives Bayes Model

1. We can see from the above results that for most of the people who were segmented as "Attack" and who died, probability of being only shot was 96% and the remainder 3% who died there was a probability of being both shot and tasered.

2. Out of all the people segmented by the police as "Attack", probability of them having signs of mental illness is 21.05%, while 78.9% of them had a probability of not having signs of mental illness.

3. Out of all the people who died and were segmented as "Attack", probability of them being male is 95.9%, while probability of them being female is 4%.

4. Out of all the people who died and were segmented as "Attack", probability of them being black is 24%, white is 46.4%, asian is 1.35%, Hispanic is 14.3%

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     attack other undetermined
##   attack          702   326           49
##   other            16    16            0
##   undetermined      0     0            0
##
## Overall Statistics
##
##                Accuracy : 0.6474
##                  95% CI : (0.6185, 0.6756)
##     No Information Rate : 0.6474
##     P-Value [Acc > NIR] : 0.5138
##
##                   Kappa : 0.027
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: attack Class: other Class: undetermined
## Sensitivity                0.97772      0.04678             0.00000
## Specificity                0.04092      0.97914             1.00000
## Pos Pred Value             0.65181      0.50000                 NaN
## Neg Pred Value             0.50000      0.69731             0.95582
## Prevalence                 0.64743      0.30839             0.04418
## Detection Rate             0.63300      0.01443             0.00000
## Detection Prevalence       0.97115      0.02885             0.00000
## Balanced Accuracy          0.50932      0.51296             0.50000
```

From the confusion matrix above, we get an accuracy of 64.74%, the reason we are getting a low accuracy is because the other and undetermined categories have many different categories unnamed reasons for deaths that are not properly mentioned by the Police and hence unsupervised models are unable to predict Other and Unsupervised categories.

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##                       A           B           H           N           O
## 0.108260185 0.016205267 0.232725636 0.162727887 0.014854828 0.008777853
##           W
## 0.456448346
```

```
##
## Conditional probabilities:
##    manner_of_death
## Y         shot shot and Tasered
##     0.95426195       0.04573805
##   A 0.88888889       0.11111111
##   B 0.94487427       0.05512573
##   H 0.94329184       0.05670816
##   N 0.98484848       0.01515152
##   O 0.87179487       0.12820513
##   W 0.95266272       0.04733728
##
##    signs_of_mental_illness
## Y       False      True
##     0.7692308 0.2307692
##   A 0.7222222 0.2777778
##   B 0.8568665 0.1431335
##   H 0.8298755 0.1701245
##   N 0.8333333 0.1666667
##   O 0.7179487 0.2820513
##   W 0.7120316 0.2879684
##
##    threat_level
## Y       attack      other undetermined
##     0.65488565 0.28482328   0.06029106
##   A 0.54166667 0.44444444   0.01388889
##   B 0.67117988 0.28336557   0.04545455
##   H 0.58506224 0.36099585   0.05394191
##   N 0.59090909 0.34848485   0.06060606
##   O 0.66666667 0.33333333   0.00000000
##   W 0.65927022 0.30226824   0.03846154
##
##    gender
## Y                          F            M
##     0.0000000000 0.0207900208 0.9792099792
##   A 0.0000000000 0.0555555556 0.9444444444
##   B 0.0000000000 0.0377176015 0.9622823985
##   H 0.0000000000 0.0262793914 0.9737206086
##   N 0.0000000000 0.0757575758 0.9242424242
##   O 0.0000000000 0.1025641026 0.8974358974
##   W 0.0004930966 0.0537475345 0.9457593688
```

Results from Naives Bayes Model 2

1. We can see from the above results that for most of the people who were Asian, 90.7% of deaths were caused by shooting, and 9.3% of deaths for asians are caused by shooting and taser shots. Similarly, 94.7% of black deaths were caused by shooting and 4.3% by shootings and taser shots. Also, white people die most of shootings with a probability of 95%

2. Out of all the people who had mental health issues, white people had the highest probability of having mental health issues. Similarly, black people have the lowest probability of having mental health issues.

3. Out of all the people who died "Other" people had the highest segmentation by police of marked as having high threat. Asian people have the lowest threat level probability.

4. Out of all the people who died, probability of being male in a race belongs to Immigrants and "other' category people had the highest probability of female deaths.

Model 3 Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Random Forest is a supervised machine learning algorithm which is based on ensemble learning.

Advantages: 1. It overcomes the problem of overfitting by averaging or combining the results of different decision trees. 2. Random forests work well for a large range of data items than a single decision tree does. 3. Random forest has less variance than single decision tree. 4. Random forests are very flexible and possess very high accuracy. 5. Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling. 6. Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

```
##
## Call:
##  randomForest(formula = threat_level ~ ., data = shooting2.train,      importance = T, ntree = 500, m
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 35.29%
## Confusion matrix:
##              attack other undetermined class.error
## attack         2829    44            0   0.0153150
## other          1326    46            0   0.9664723
## undetermined    194     4            0   1.0000000
```

Our Random Forest Model gives an out of box estimate of error as 35.18%

As discussed above, our model is unable to segment between other and undetermined categories as there is not a proper identifier for these classes, while Attack class has only a 1.3% error.

```
##               Actual
## Predicted      attack other undetermined
##   attack          707   328           49
##   other            11    14            0
##   undetermined      0     0            0
```

```
## [1] 0.6501353
```

From the confusion matrix for random forest, we get accuracy of 64.65%.

Conclusion

1. Black people, Native Americans and Hispanic people are the most affected races.
2. California is the most affected state with the highest number of deaths.
3. There is a higher rate of deaths during weekdays and during the first 3 months of the year.
4. There is evidence for police brutality based on race and hence we reject the null hypothesis that said there is no police brutality difference based on race.

5. The models run are not very accurate with determining the Threat Level. There is a rough 35% error across all models run.
6. The best model out of the 3 is Naives Bayes Model. It had the best accuracy and most correctly identified true positives for attack class.

References

1. Dataset from Kaggle: https://www.kaggle.com/jpmiller/police-violence-in-the-us
2. Dataset1 Source: Mapping Police Violence, https://mappingpoliceviolence.org/aboutthedata
3. Dataset2 Source: The Washington Post, https://github.com/washingtonpost/data-police-shootings