Exploring Fatal Police Shootings in the United States

Kidus Asfaw, Laura Niss, Zoe Rehnberg, James Wu

STATS 601 Final Project
Winter 2017

Contents

1 Introduction

1.1 Washington Post Fatal Police Shooting Data

The Washington Post aims to record every fatal police shooting by an on duty police officer starting January 1, 2015. As stated on their website, the data is collected "by culling local news reports, law enforcement websites and social media and by monitoring independent databases such as Killed by Police and Fatal Encounters. The Post conducted additional reporting in many cases" [6]. As of this writing, the dataset contains 2233 fatal shootings and tracks 13 variables.

1.1.1 Data Challenges

This dataset consists almost entirely of categorical variables. While not inherently a problem, this structure makes it more difficult to work with some of the variables. For example, both city and state are factors, which means they give almost no information about the spatial relationship between observations. To deal with this, we converted the city and state variables into latitude and longitude, which allows us to better capture spatial effects. Additionally, victims were carrying 66 different weapons when they were shot, most of which appeared in only a handful of incidences. To make weapon information more usable, we consolidated these 66 types of weapons into 7 categories, based on the 6 most common weapon types. The 7 categories are gun, knife, vehicle, undetermined weapon, toy weapon, unarmed, and other.

1.2 Motivating Questions

1.2.1 Predicting Race

When considering the data, we were curious to see if knowing the circumstances of a fatal police shooting could help us predict the race of the victim. If the variables in our dataset help us to predict race, this could be used as evidence of a systemic problem of disproportionate use of force against certain races.

Our interest in predicting race was further motivated by some interesting patterns we saw in the data, as seen in Figures 1, 2, and 3. For instance, about 25% of people fatally shot by police since 2015 have been black, which is more than double the proportion of black Americans (12.2%). On the other

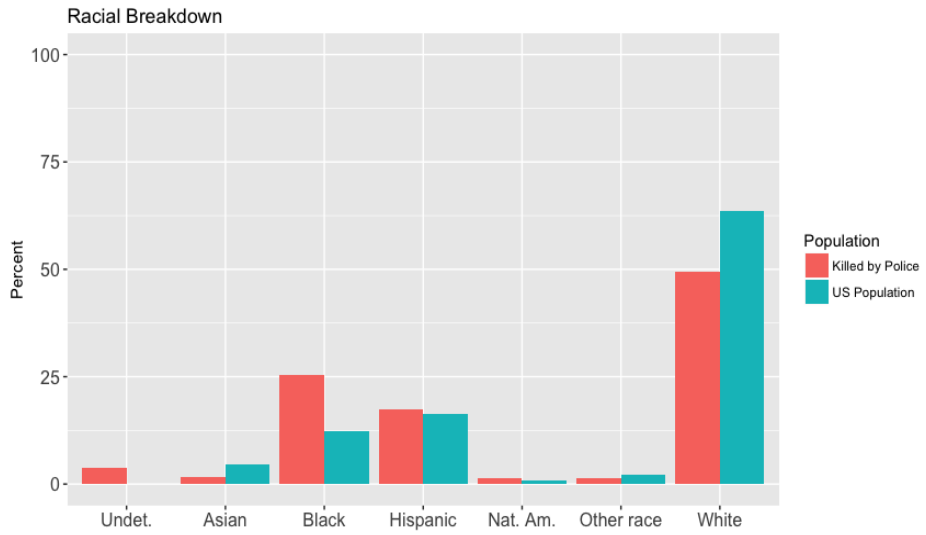


Figure 1

hand, just under 50% of those killed by police were white, while 63.7% of the U.S. population is white.

Similarly, there were interesting racial trends between the levels of some variables in the dataset. For instance, we found differences in the racial break-

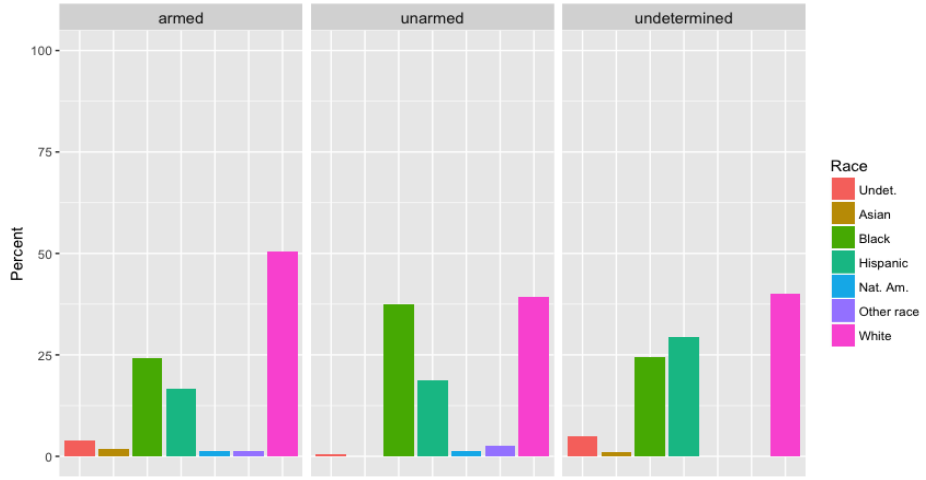


Figure 2

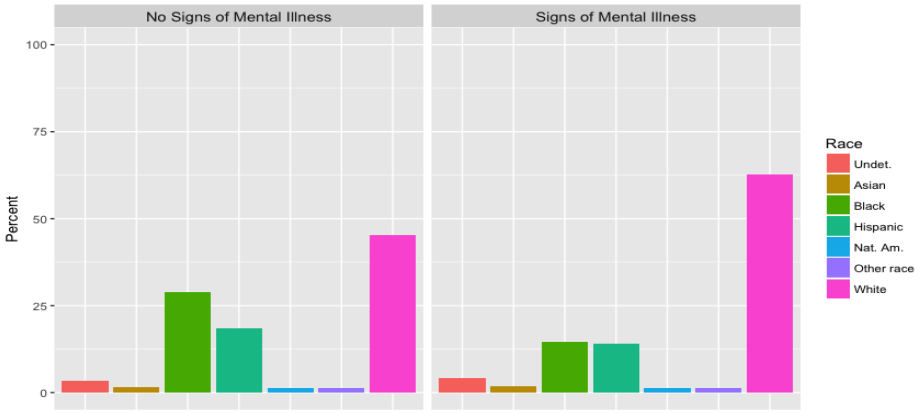


Figure 3

down among armed and unarmed victims and among victims who did and did not show signs of mental illness. These patterns reinforced our belief that race is an important factor in fatal police shootings.

Because there has been increased attention on fatal police shootings among black communities and there were so few victims from other racial backgrounds, we decided to consolidate the racial categories for our analyses. We defined race as "Black," "White," and "Other" during cluster analysis and race prediction.

1.2.2 Predicting Body Cameras

Another question of interest was whether we could classify the use of body cameras using the other covariates. A major debate in police accountability is the efficacy of body cameras in preventing the excessive use of force by police. If we accept the premise that body cameras do not make a difference, then we would be unable to predict either class at a rate better than a coin flip. However, if we were able to predict whether police officers were wearing body cameras, this might indicate, for example, that police officers behave differently when wearing body cameras. Alternatively, they might behave the same but report the specifics of an incident less accurately when they do not wear body cameras. Finally, different cities and states have different rules regarding the use of body cameras. Thus, we may classify based on location rather than on differences in reporting or behavior.

2 Data Visualization

To visualize the whole dataset, we needed to implement a form of dimension reduction. However, due to the categorical nature of the data, common techniques like principle component analysis and factor analysis do not make sense – our data would have violated the model assumptions. It is possible, however, to construct a distance matrix based on categorical variables, which allows us to use classical multidimensional scaling (CMDS).

2.1 Gower Distance

Because the data are largely made up of categorical variables, using Euclidean distance to create a distance matrix does not make sense. Instead, we used Gower distance, which is a measure that defines the distance between two observations based on both continuous and categorical variables [4]. The distance between observations i and j is defined as follows:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_k \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} w_k \delta_{ij}^{(k)}} \tag{1}$$

where w_k and $\delta_{ij}^{(k)}$ are weights for the k^{th} variable. For categorical variables,

$$d_{ij}^{(k)} = \begin{cases} 0 & \text{if } x_{ik} = x_{jk} \\ 1 & \text{if } x_{ik} \neq x_{jk} \end{cases}$$

and for continuous variables,

$$d_{ij}^{(k)} = \frac{|x_{ik} - x_{jk}|}{\text{range}(x_{\cdot k})}$$

For the purposes of our exploratory data analysis, we set the weight terms to be 1. In this setting, the Gower distance reduces to the average of the distance between each variable.

To validate our use of this distance measure, we looked at the individuals who were indicated to be most similar and most dissimilar by the Gower distance. In Figure 4 (a), the individuals are (almost) exactly the same on every covariate, and in Figure 4 (b), the individuals have the same values for none of the covariates. These two observations provide some validation of the use of Gower distance in this application.

	id	manner_of_death	armed	age	gender	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority
2015	2238	shot	gun	33	M	False	other	Not fleeing	False	-118.26	33.81	other
1914	2134	shot	gun	33	M	False	other	Not fleeing	False	-118.25	33.97	other

(a)

	id	manner_of_death	armed	age	gender	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority
2124	2372	shot	knife	30	F	True	other	Car	False	-97.77	30.33	black
1637	1829	shot and Tasered	gun	28	M	False	attack	Foot	True	-149.34	67.09	other

(b)

Figure 4

2.2 Classical Multidimensional Scaling

Using the distance matrix constructed based on the Gower distance, we implemented CMDS as a way to visualize the complete dataset in two dimensions. Additionally, we colored the observations based on the levels of various variables to look for patterns in the data. We found no clear trends based

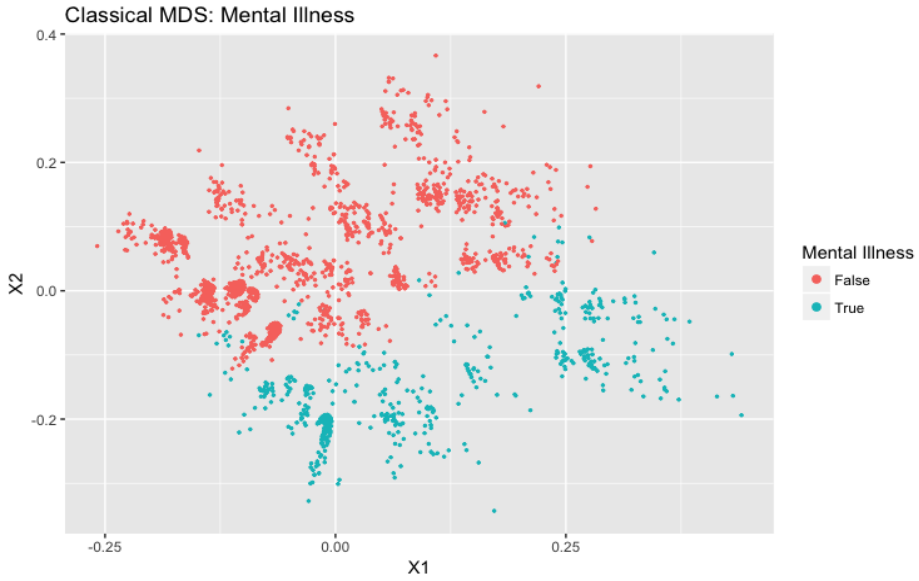


Figure 5

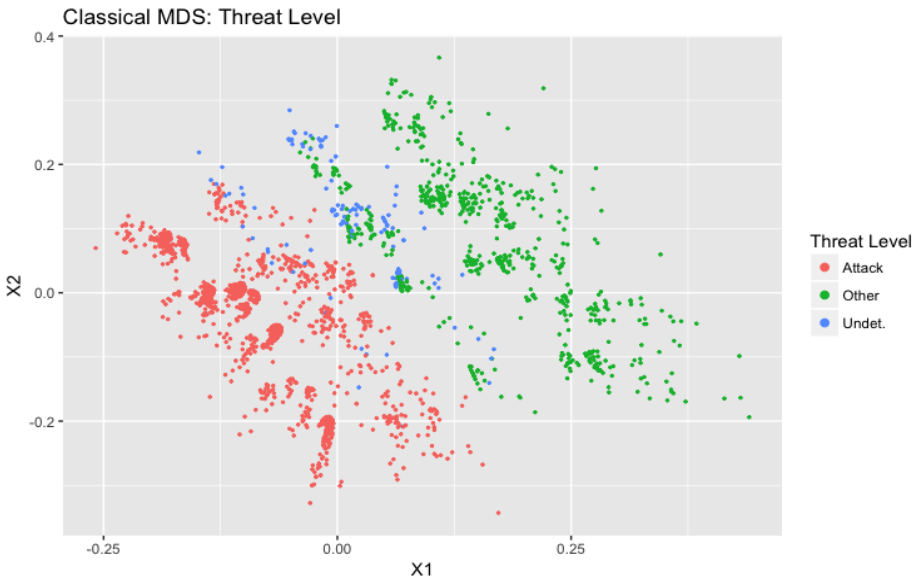


Figure 6

on race or body camera (not pictured), but did find interesting trends in the mental illness status and threat level of the victims (see Figures 5 and 6).

3 Clustering

3.1 K-medoids

To explore natural clustering of the data, we chose to use the k-medoids algorithm. This method is similar to k-means, but instead of using within-group sample averages, k-medoids uses data points $C_k \in (X_1, \ldots, X_n)$ as the centers for each group $1, \ldots, k$. This method of clustering is particularly interpretable because the returned centers are the data points that best exemplify the points in each cluster. Let $C(i)$ be the current classification of point i. Then each center C_k is usually chosen to minimize the Euclidean distance

$$\sum_{i, C(i)=k} ||X_i - C_k||^2 \tag{2}$$

However, since the distance measure we are using is the Gower distance, we

instead choose C_k to minimize

$$\sum_{i,C(i)=k} d_{ik} \tag{3}$$

To find the centers, we use the k-medoids algorithm defined below [3]:

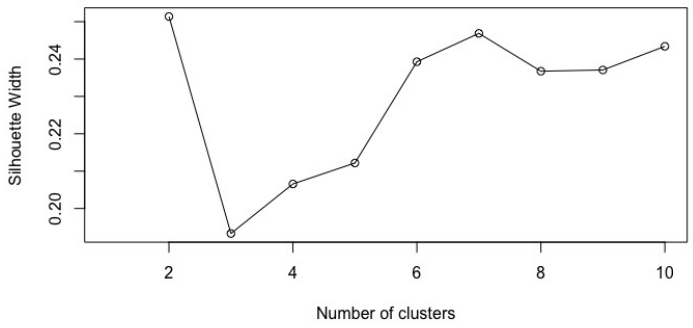
1) Initialize $C_1, \ldots, C_k$ as random points from the data.
2) Find the cluster center C_k closest to X_i and set $C(i) = k$.
3) Find new C_k such that it minimizes $\sum_{i,C(i)=k} d_{ik}$.
4) Repeat until $\sum_{i,C(i)=k} d_{ik}$ converges for each k.

To determine the number of clusters in the data, we examined the silhouette. The silhouette is defined as

$$\frac{1}{n}\sum_{i}^{n} \frac{b(i) - a(i)}{max\{a(i), b(i)\}} = \frac{1}{n}\sum_{i}^{n} s(i) \tag{4}$$

where $a(i)$ is the average dissimilarity between i and points in its cluster, and $b(i)$ is the average dissimilarity between i and points in the closest neighboring cluster. For each i, $-1 \leq s(i) \leq 1$, with larger numbers suggesting better clustering [5].

We ran k-medoids for $k = 2, \ldots, 10$, with the resulting silhouettes:



Since $k = 2$ has the largest silhouette, and the next largest, $k = 7$, would be less interpretable, we looked only at the results for two clusters.

As shown in Figures 7 and 8, the two clusters differ mainly on the variables `armed`, `threat_level`, `lon`, and `minority`. The longitude and latitude for cluster 1 is around Arkansas, and for cluster 2 is around Arizona.

manner_of_death	armed	age	gender	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority	cluster
shot :1530	gun :1144	Min. : 6.00	: 0	False:1241	attack :1334	: 25	False:1451	Min. :-168.02	Min. :19.56	black:430	Min. :1
shot and Tasered: 79	unarmed : 90	1st Qu.:27.00	F: 64	True : 368	other : 192	Car : 247	True : 158	1st Qu.:-105.04	1st Qu.:33.52	other:251	1st Qu.:1
	vehicle : 84	Median :35.00	M:1545		undetermined: 83	Foot : 195		Median : -90.44	Median :36.31	white:928	Median :1
	toy weapon : 72	Mean :37.05				Not fleeing:1075		Mean : -94.49	Mean :36.86		Mean :1
	knife : 71	3rd Qu.:46.00				Other : 67		3rd Qu.: -82.42	3rd Qu.:40.08		3rd Qu.:1
	undetermined: 54	Max. :91.00						Max. : -68.10	Max. :71.29		Max. :1
	(Other) : 94										

Figure 7: Cluster 1

manner_of_death	armed	age	gender	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority	cluster
shot :496	knife :254	Min. :13.00	: 0	False:399	attack : 69	: 20	False:491	Min. :-168.02	Min. :19.56	black:123	Min. :2
shot and Tasered: 75	unarmed : 70	1st Qu.:26.00	F: 28	True :172	other :456	Car : 80	True : 80	1st Qu.:-118.25	1st Qu.:33.53	other:297	1st Qu.:2
	vehicle : 62	Median :33.00	M:543		undetermined: 46	Foot : 62		Median :-110.97	Median :35.49	white:151	Median :2
	gun : 45	Mean :35.01				Not fleeing:394		Mean :-103.99	Mean :36.37		Mean :2
	undetermined: 43	3rd Qu.:42.00				Other : 15		3rd Qu.: -86.00	3rd Qu.:39.53		3rd Qu.:2
	toy weapon : 23	Max. :86.00						Max. : -68.01	Max. :61.58		Max. :2
	(Other) : 74										

Figure 8: Cluster 2

These differences are also reflected in the medoids for cluster 1 and 2, respectively:

	id	manner_of_death	armed	age	gender	race	city.state	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority
2113	2360	shot	gun	33	M	W	Bull Shoals AR	False	attack	Not fleeing	False	-92.58	36.38	white

	id	manner_of_death	armed	age	gender	race	city.state	signs_of_mental_illness	threat_level	flee	body_camera	lon	lat	minority
1404	1591	shot	knife	34	M	N	Parker AZ	False	other	Not fleeing	False	-114.29	34.15	other

4 Classification

4.1 Predicting Race

Based on our data exploration, we are first interested in predicting the race of the victim based on the circumstances of the fatal shooting. Specifically, we are interested in classifying the race of the victim as "Black", "White" or "Other" (`B`, `W` and `O`). As baseline predictive models, we will apply multi-class support vector machine (SVM) and multinomial logistic regression on all the variables. We will then try two additional methods to improve the baseline models – stratifying by region and feature selection.

4.1.1 Baseline Predictive Models

Recall the formulation of classical SVM as a margin maximization problem in a two-class setting. Because we have three classes, we will need a multi-class variant of this method. In addition, because most of our variables are factors, we will need to encode these as dummy variables to implement SVM.

The most common extension of classical SVM to a multi-class scenario uses the so-called *one-versus-all* approach. Assuming we have a K-class problem, we construct K classical SVM models where the k^{th} model uses observations belonging to class k as positive training examples and observations belonging to all other classes as negative training examples. We then predict using:

$$y(x) = \max_k y_k(x) \tag{5}$$

We can also extend logistic regression to a multi-class setting. By maximizing the conditional probability and using iteratively reweighted least squares (IRLS), we can find solutions to the coefficients for our explanatory variables. The results of predicting race using all variables is shown in the table below.

	Test Set Error	Training Set Error
Linear SVM	0.405	0.417
Polynomial SVM	0.405	0.393
Gaussian SVM	0.393	0.385
Logistic Regression	0.405	0.427

4.1.2 Stratifying by Region

To improve upon the error rates from the baseline model, we wanted to see if splitting our data by an informative variable and training separate models for each subset of the data could lead to better predictions. Specifically, we wanted to train separate models for each region of the U.S. and examine the resulting error rates. The results for SVM are in the table below.

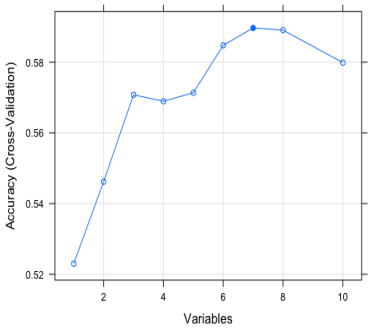
	Test Set Error	Training Set Error
Northeast	0.439	0.276
Midwest	0.374	0.059
South	0.409	0.169
West	0.471	0.188
Pooled	0.393	0.385

As we can see, stratifying our data by region results in better overall training error rates than our baseline pooled-data classifier. However, it only improves testing accuracy rates for the Midwest region. This could be because the three other regions do not have as much of a "region effect" and, in fact, suffer from not having the more general training data that could be gained from the other regions.

4.1.3 Feature Selection and Combination of Methods

Feature selection is used to produce smaller models that are more interpretable, less prone to the curse of dimensionality and, in some cases, more accurate. We used recursive feature elimination (RFE) to start with a full SVM model and iteratively eliminate the least important variables to ultimately produce a better classifier than our baseline model.

As shown in Figure 9, we can improve over the baseline model by excluding 3 variables: `flee`, `manner_of_death` and `body_camera`. The training set error rate from classifying using our 7-variable model without splitting by region is 0.375 while the test error rate is 0.398. As we had hoped, we got a performance that is equivalent to the full model using three fewer variables. From the RFE output, we can see that longitude and latitude (both of which are location parameters) are very important features. With this in mind, we added an interaction term between longitude and latitude to our 7-variable model. The training error in this case was 0.321 while the test error rate was




```
Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.5230 0.2462    0.02906 0.05262
         2   0.5462 0.2605    0.04939 0.07722
         3   0.5708 0.2954    0.02759 0.04567
         4   0.5690 0.2946    0.04643 0.07543
         5   0.5713 0.2939    0.03917 0.06454
         6   0.5848 0.3130    0.02501 0.04204
         7   0.5897 0.3165    0.02767 0.04469        *
         8   0.5891 0.3132    0.02924 0.04623
        10   0.5799 0.3053    0.03477 0.05571

The top 5 variables (out of 7):
   lon, age, lat, signs_of_mental_illness, weapon_cat
```

Figure 9

0.427. This improvement on the training set and worsening on the test set seems to indicate that the interaction term introduced overfitting into our model. Finally, in the table below, we show the pooled and region-specific results using only the 7 top variables. As we would expect, the training error rates are worse in general, but the test error rates are comparable to the full model case stratified by region.

	Test Set Error	Training Set Error
Northeast	0.390	0.276
Midwest	0.462	0.118
South	0.44	0.236
West	0.429	0.271
Pooled	0.398	0.375

4.2 Predicting Body Cameras

In this section, we attempt to classify observations as body camera ("BC" or "True") or no body camera ("No BC" or "False"). We use the random forest algorithm [1] as our classifier for various reasons. Using a tree-based classifier allows the variables to interact without explicitly specifying the interaction terms. Although random forests are less interpretable than a single decision tree, they still provide variable importances. This is helpful, as we wish to infer which variables are important in classifying BC and No BC. Unlike a single decision tree, we are also able to estimate the conditional probability

that an observation is BC by calculating the proportion of votes for BC among the trees constructed. Finally, we also hope to take advantage of the out-of-bag ("OOB") observations to validate our models.

4.2.1 Unbalanced Classification

One major issue in predicting BC vs. No BC is that the classes are very unbalanced. In the training set, roughly 90% of the observations are No BC. Furthermore, based off our initial exploration of the data, we expect that the classes are not well separated. Thus, our initial attempts yield a classifier in which nearly all the observations are classified as No BC. While this results in an overall error rate of just 11%, the error rate for the BC class is almost 100%, as we can see in Figure 10.

```
Call:
 randomForest(x = X.train, y = Y.train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 11.09%
Confusion matrix:
      False True class.error
False  1489    4 0.002679169
True    182    2 0.989130435
```

Figure 10

Because we wish to predict both classes equally well, the overall error rate is less important than the error rates for the individual classes. To this end, rather than using $0-1$ loss, we assess our results using the following loss function:

$$\mathrm{L}(y_i, \hat{f}(x_i)) = \begin{cases} 0, & y_i = \hat{f}(x_i) \\ 1, & y_i = 0; \hat{f}(x_i) = 1 \\ c, & y_i = 1; \hat{f}(x_i) = 0 \end{cases} \tag{6}$$

where c is some real number, $y_i = 1$ for class BC, and $y_i = 0$ for class No BC. When $c = 1$, this is simply $0-1$ loss. For our classification problem, we use $c = 9$, as the ratio of No BC to BC observations is roughly 9-to-1. This gives equal weight to the classes overall (*i.e.*, the same prediction error in each class results in the same loss for each class as a whole).

4.2.2 Balanced Sampling

We used two sampling schemes to obtain a balanced training set. First, we oversampled the BC class in our training set: we kept all observations in the No BC class and then sampled from the BC class with replacement until we had the same number of observations of both classes. We then fit a random forest to the oversampled training set. The result is shown in Figure 11.

```
Call:
 randomForest(x = X.up, y = Y.up)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 1.24%
Confusion matrix:
      False True class.error
False  1456   37  0.02478232
True      0 1493  0.00000000
```

Figure 11

It initially appears as though the oversampling approach fixes the issue described earlier, as the OOB error rates are quite low for both classes. However, the cross validation (CV) error rate for the BC class is still over 90%. One possible explanation is that even though we obtain a balanced training set using this approach, most of the observations in the BC class are replicates of each other. Regardless, it appears that the classifier overfits to the training data.

We next tried undersampling the No BC class: we kept all observations in the BC class, and then sampled without replacement from the No BC class until we had the same number of observations in the BC and No BC classes. While the OOB error rates are higher with this method, the CV error rate for BC is lower (roughly 45%). In both classes, the OOB error rates are slightly better than a coin flip. Undersampling does not seem to suffer from the same overfitting issues as oversampling; however, it does not take advantage of all the available information as we leave out most of the No BC class.

4.2.3 Using a Different Probability Cut-Off

In this section, we try to improve performance with the regular and oversampled training sets. To accomplish this, we used a different conditional

```
Call:
 randomForest(x = X.down, y = Y.down)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 44.23%
Confusion matrix:
      False True class.error
False   107   75   0.4120879
True     86   96   0.4725275
```

Figure 12

probability to classify an observation as BC or No BC. Normally, we would use the classifier

$$\hat{f}(x_i) = \mathbb{I}[\hat{P}(Y_i = 1|X_i) > 0.5] \tag{7}$$

where $\hat{P}(Y_i = 1|X_i)$ is the estimated conditional probability from the random forest. However, if we instead use

$$\hat{f}(x_i) = \mathbb{I}[\hat{P}(Y_i = 1|X_i) > k] \tag{8}$$

for some $k \in (0, 0.5)$, we predict BC even when we are less certain. Thus, we are essentially making it more expensive to misclassify a BC observation as No BC.

To evaluate this method, we fit a random forest model to each training set (regular, oversampled, and undersampled) and selected k using 5-fold cross validation. To avoid the overfitting issue we observed in the oversampling case, we performed the sampling after splitting our training set into folds. That is, we split our training data into five folds and, after leaving each fold out, construct the over or undersampled training set from the remaining four folds. We then fit our model to that data set. We get $k = 0.10$, 0.17, and 0.5 for the regular, oversampled, and undersampled training sets, respectively.

4.3 Results

In the table below, we examine the training and test set errors split by class. In all three training sets, the random forest overfits to the training data, as shown by the low training set error rates. Looking at the test set error, the regular training set appears to outperform the undersampled training set.

This may occur because the regular training set makes use of all the observations in the training set. However, we would also expect the oversampled training set to outperform the undersampled training set for similar reasons, but we see that the error rates are very similar for both training sets.

	Test Set Error		Training Set Error	
	BC	No BC	BC	No BC
Regular	0.3673	0.4523	0	0.0847
Oversampling	0.4490	0.4257	0	0.1176
Undersampling	0.4286	0.4568	0	0

As noted earlier, we not only wish to classify BC and No BC, but also to determine in what way the two classes differed. According to the random forest results, the most important variables were latitude, longitude, age, race, and weapon category. The location variables were by far the most important variables. It therefore seems that many of the differences between the two classes can be explained by location.

5 Conclusions

In our exploratory analysis, we observed some differences among racial groups, including a disproportionately large number of unarmed black victims. We aimed to quantify these difference by predicting race from the circumstances of the shooting.

This task turned out to be quite difficult. Both SVM and logistic regression gave similar error rates of about 40%. This is notably better than guessing at random, but does not fully explain differences seen in the data. Nevertheless, there do appear to be differences in the circumstances surrounding fatal police shootings of white, black, and other (Hispanic, Asian, Native American, and other) victims.

We had similar difficulty when predicting whether an officer was wearing a body camera or not. Because latitude and longitude were the two most important variables in the classification, it is possible that the classifier is simply finding differences in state body camera laws. It is also possible that adding instances from non-fatal police shootings could give more insight into circumstantial differences.

References

[1] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[2] Demography of the United States. (2017, April 10). In *Wikipedia, The Free Encyclopedia*. Retrieved April 11, 2017.

[3] Kaufman, L., and Rousseeuw, P. J. (1987). Clustering by means of medoids. in Y. Dodge (editor) *Statistical Data Analysis based on L1 Norm*, 405-416.

[4] "Package 'cluster'." The Comprehensive R Archive Network, 16 Mar. 2017. Web. 9 Apr. 2017. https://cran.r-project.org/web/packages/cluster/cluster.pdf.

[5] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

[6] Washington Post (2017). *Police Shootings Database*. Retrieved from https://github.com/washingtonpost/data-police-shootings, April 9, 2017.

Code available at: https://github.com/laurakn/Washington-Post-Fatal-Police-Shootings