CPDBP: Injury Trends
Checkpoint 4: Machine Learning

By: Hawkins Gay, Alex Leidner, Ramsey Wehbe

Data Science Seminar Fall 2020

Introduction – Data Processing:

Given that this section's questions have similar variables of interest from the database, we will discuss our data processing pipeline here as it applies to both questions. The main difference will be which category will be labels for each question which will be discussed separately.

First we collected the features of interest. From trr_trr we colleceted, id, subject_injured, subject_alleged_injury, district_name, lighting_condition, indoor_or_outdoor, weather_condition, officer_in_uniform, officer_injured, officer_rank, subject_age, subject_gender, subject_race, officer_id, trr_datetime. We joined this data by the correlating officer id with data_officer to get the officers: gender, race, age (year from date time – birth year). We did feel that the environmental conditions and lighting may provide meaningful information as to whether or not an injury occurred so these were included in our model despite not being part of our question to avoid ignoring a confounder.

We joined this information about the officers and subjects using the trr_id with trr_actionresponse. This table contains multiple types of force used in each use of force report. To avoid duplicating use of force reports we extracted each type of force into its own column. The types of force are firearm, taser/impact_weapon, chemical weapon, other_weapon, member presence, and verbal commands.

Next we created dummy variables for all of the categorical features in the dataset in a similar fashion to weapon type so that these categorical features could be used as input to ML classifiers as recommended in the Scki-kit Learn documentation.

Then we imputed the Null variables from the 67019 use of force reports. The missing variables were: 3638 lighting conditions, 3605 indoor/outdoor, 3799 weather conditions, 2005 district name, 878 subject race. The remaining columns had less than 300 missing values.
We used scikit-learns iterative imputer to impute values between 0 and 1 for the missing categories.

After this we set aside 20% of our data as test data not to be used we finalized our model.

Both questions we used a random forest as one of our models. Class weighting was used for both questions given class imbalance with underrepresentation of the positive class (injury, and lethal force). We used two different approaches to training a model for these questions to demonstrate the different approaches to hyperparameter tuning:

1) For question 1, we manually tuned hyperparameters using a hold-out validation set. Additionally, given poor performance in a RF model for question one we further trained a penalized logistic regression model and K-nearest neighbors.

2) For question 2, we used an automated method of hyperparameter grid search to find the optimal hyperparameters using Sci-kit learn's GridCVSearch with 5-fold cross-validation using the following hyperparameter search space:

```python
clf = CalibratedClassifierCV(base_estimator=RandomForestClassifier(random_state=42,
                                                                  class_weight = 'balanced',
                                                                  n_jobs = 48))
param_grid = {'base_estimator__max_depth': [10, 25, 50, None],
              'base_estimator__max_samples': [1000, 10000, 30000, None],
              'base_estimator__max_features': [1, 25, 50, 'auto', 'log2', None]}
```

```python
search = GridSearchCV(clf, param_grid, cv=2, n_jobs = 48, scoring = 'roc_auc')
```

After comparing 72 models the best ROC Curve was found for a max depth of 25 and None for max_samples(use all samples at each split) and max_features(use all features as opposed to sqrt(n_features)).

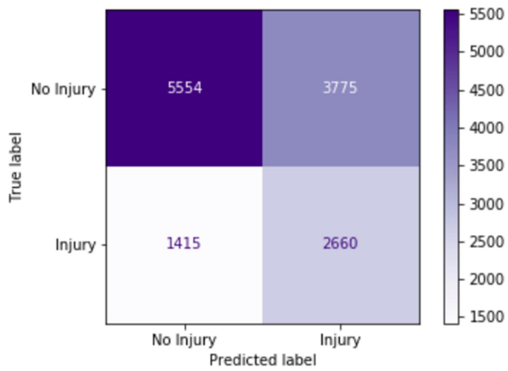
We calculated performance metrics for all models including overall accuracy, precision, recall, F1 score, and area under the receiver operator curve (AUC). Given AUC is a prevalence independent metric that focuses on pure discriminative performance of the model, we used AUC as a metric to optimize our models. Finally, for both questions we computed the Youden index (sensitivity+specificity -1) of the corresponding ROC curve to find the optimal decision threshold for discrimination between positive and negative classes, and recalculated class predictions based off of this threshold.

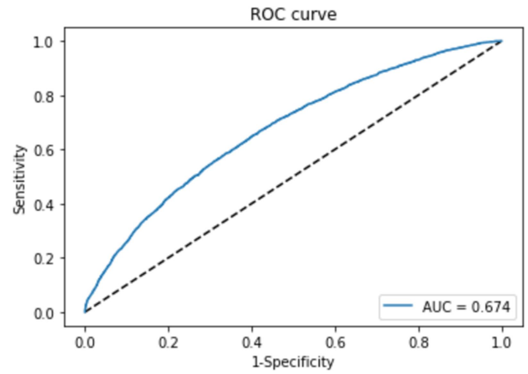
Question 1: Predict police encounters (based on citizen and police demographics, alleged crime, neighborhood) where use of force incidents are likely to result in injury.

For our first question, we were interested in predicting which use of force events that would result in injury. Injury included true subject_injured from the trr_trr data table as well as subject_alleged_injury.

The first model we assessed was a random forest model. We used the validation set to manually fine tune the hyperparameters. For random forest, the most accurate set of hyperparameters was [n_estimators=100, class_weight='balanced', max_depth=25]. We also compared the random forest classifier to a kNN classifier and a logistic regression classifier. The best predictions ended being completed with the random forest model, and we were able to increase the predictive accuracy of our model by utilizing the AUROC values in the predictions. This is the model we ended up being used on the testing data

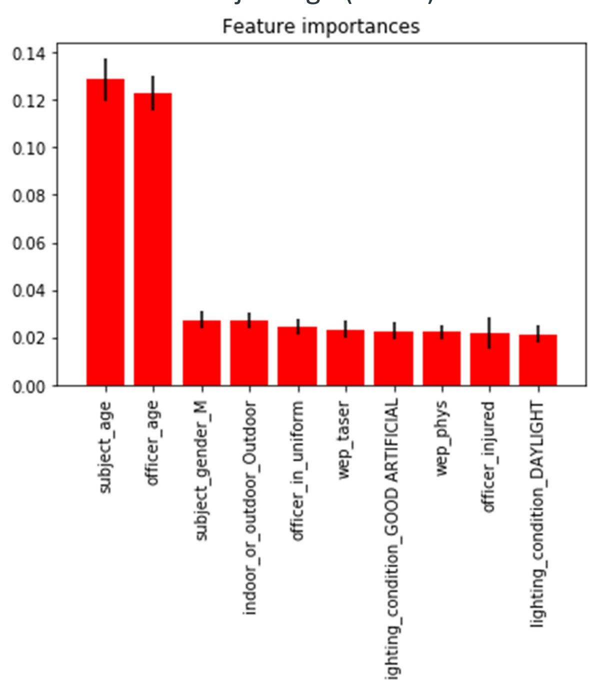
When applying the random forest model to the testing data we returned the following results :





	precision	recall	f1-score	support
0.0	0.80	0.60	0.68	9329
1.0	0.41	0.65	0.51	4075
accuracy			0.61	13404
macro avg	0.61	0.62	0.59	13404
weighted avg	0.68	0.61	0.63	13404

There were more false positives then true positives, which means that the model was very sensitive for predicting injury, though they may be misclassified. However, the accuracy and recall of the model were still greater than 60%. By examining the AUROC you get a better sense of the models true discriminative performance; the black dotted line represents a classifier that is no better than a coin toss and our model was slightly better than this, suggesting it is not that successful of a classifier. Unfortunately, it doesn't appear that there are any strong predictors of injury in this dataset based on or modeling. Of the data that we did assess, the strongest features were subject age (0.129) and officer age (0.122).

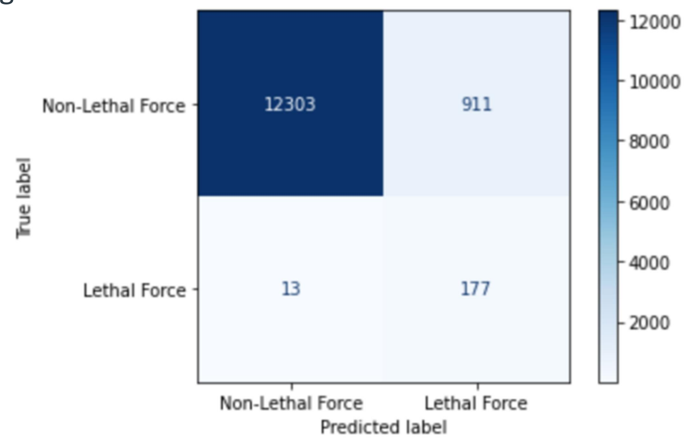


Question 2: Predict police encounters more likely to result in lethal uses of force vs. less-than-lethal uses of force.

For this question we were interested in predicting the factors that could result in a police officer deciding to use lethal force. We interpreted this target as an officer using their firearm, as we felt this is a different event and decision than an officer pushing, yelling, or using a Taser. In order to do this we removed subject injury and used firearm use as a target for our Random Forest model.

Initially we had a very poor prediction value because random forest requires a vote of 0.5 probability by default or higher in order to choose a class. This was very difficult to reach given that only about 1000 events out of the 67,000 data points contained firearm use. While this maximized accuracy, this was not useful because a model that said there would never be any firearm use would have greater than 98% accuracy. We did not feel this helped us predict our outcome of interest. To balance this, we chose to maximize our model based upon area under the ROC curve. Our grid CV search created a model as specified in the introduction. This model for this data had an AUC score was 0.9738 with 5 fold cross validation on the training data. Next we were able to apply our model to our test data. Figure 2.1 shows the confusion matrix that we achieved as well as other characteristics. Figure 2.2 shows the ROC curve on the test data. While there are many false positives we felt this to be preferable than to frequently miss deadly weapon use. We were able to limit the false negatives as well as provide a very high true negative rate and maintain accuracy at 0.93. Finally we can see that the AUC curve maintains its shape and predictive values over the test data so that we did not over-fit our data. This is probably because we limited the depth of the tree and used cross validation to do our best to prove that this model would generalize.

Figure 2.1: Random Forest Confusion Matrix

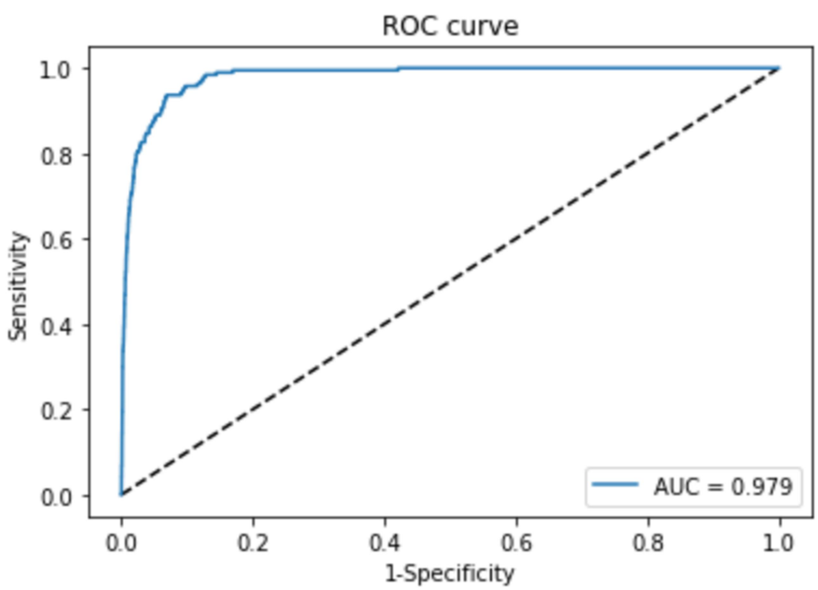



```
Classification Report
                  precision    recall  f1-score   support

Non-Lethal Force       1.00      0.93      0.96     13214
    Lethal Force       0.16      0.93      0.28       190

        accuracy                           0.93     13404
       macro avg       0.58      0.93      0.62     13404
    weighted avg       0.99      0.93      0.95     13404
```

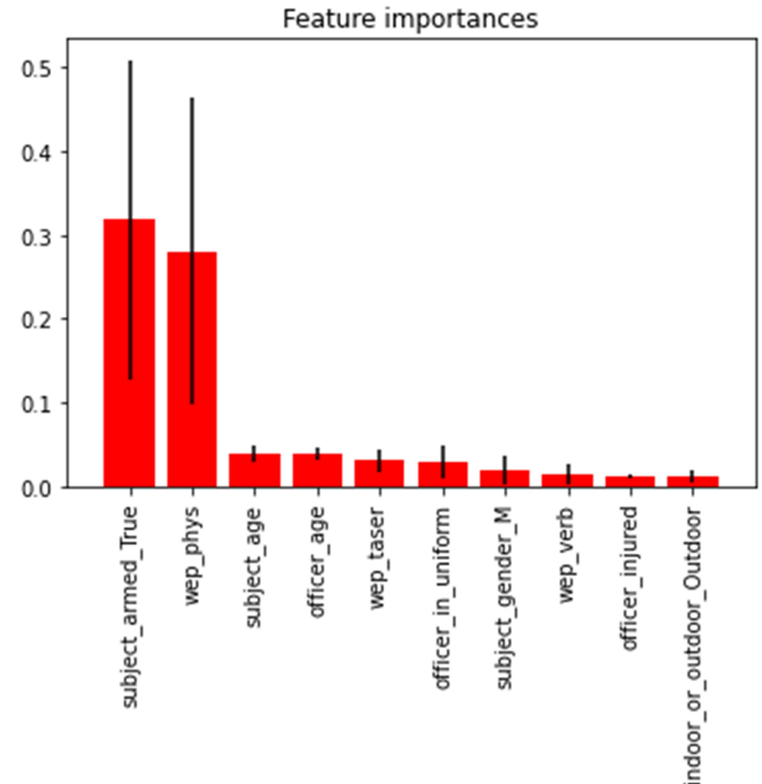
Figure 2.2: Random Forest ROC Curve



The feature importance that this forest possesses is displayed in figure 2.3. To no surprise one of the most important features that is correlated with an officer using lethal force is if their subject is armed. The officer having to use physical force during the encounter was also correlated. Other factors such as the subject's age and officer age and gender of the subject also provided some predictive weight.

While officer's brandishing their firearm in response to a subject being armed does seem obvious, the use of physical force being a high risk factor for escalation to a firearm is very interesting especially as use of physical force was quite common throughout the data. This could lead to further officer training that just because physical force was used a reminder could be implemented this does not mean lethal force has to follow.

Figure 2.3: Random Forest Feature Importance(Top 10)



Conclusion:

Our initial hypothesis of demographics resulting in subject injury did not appear to produce very strong models. Throughout a random forest, KNN, and logistic regression we could not make a model that could predict much better than a coin flip. This is somewhat expected as our previous checkpoint analysis has shown that while Blacks are on the receiving end of more force, that injuries and types of weapons used are fairly consistent throughout these demographic variables occurring a third of the time, making their stratifications difficult to correlate into a model.

Furthermore our second question's model does indicate a correlation between use of physical force and lethal force was predictive in our random forest model. While there are some times that officers must brandish their gun in response to armed subjects there is also the potential for scrutiny of gun use as an escalation of physical force. This could serve as a starting point for further research into officers that are frequently involved in physical conflict, as they may be a higher risk for using lethal force in the future.