Final Report: Predicting Civilian-Officer Outcomes

Lucy Du (dd483), Elina Hvirtsman (eh582), Julia Ng (jen67)
December 13, 2020

1 Problem Statement

Our project's goal is to understand what factors most influence civilian and police interaction outcomes and determine if conduct by the Minneapolis police precincts can be predicted. Given demographic and situational factors, we fit and develop models that predict whether the civilian will be issued a citation in the event of being stopped by police, or which police force the civilian will experience in more severe situations.

2 Dataset

The two main datasets are sourced from Open Minneapolis. The goal of the Minneapolis Open Data Portal is to encourage access to data managed by the City of Minneapolis, making content available to be freely used, modified, and shared by anyone for any purpose. The datasets are refreshed on a daily basis by 9:30 AM. For the purpose of this project, we use the finalized dataset published on October 24, 2020.

The Police Stop dataset spans from July 6, 2017 to October 24, 2020, and the Police Use of Force dataset spans from January 1, 1970 to October 23, 2020. Both provide basic information about the civilian like race and gender, as well as information about each civilian-officer interaction like if there was a search of the person, the primary offense, and the specific problem, such as suspicious person, traffic law enforcement, attempt pick-up, etc. Police precinct, neighborhood, and latitudinal/longitudinal data is present too. Our outcome variable of interest is citation issued for Police Stop, which contains either yes or no. Our outcome variable of interest is force type for Police Use of Force, which contains values like bodily force, taser, chemical irritant, etc. There are 12 features and 116,165 examples in the first dataset, and 20 features and 29,787 examples in the second dataset.

3 Data Preprocessing

To make our datasets fit our problem statement, we dropped any observations with missing citationIssued or ForceType values. To make our dataset more approachable, we converted some nominal values into one-hot encoding. For example, the single Race column, which contains values like Asian, Black, East African, Latino, Native American, Other, Unknown, and White, was converted into eight columns with a 1 in the applicable column and a 0 in the others. For variables like Neighborhood, which have 86 possible entries, we keep the values for the top 5 most common and convert the remaining into a category Other. This column is then converted again through one-hot encoding. Finally, the aggregated Date column is separated into more usable Year, Month, Day, Hour, and Minute columns.

The datasets are comprehensive and complete with intuitive values. Histograms were created for the numeric variables like latitude and longitude to check for outliers. In the Police Stop data, there are around 700 observations with missing latitude and longitude, which are dropped given it is less than 1% of our entire dataset. Frequency tables were created for the nominal variables like race, problem, and neighborhood to check for outliers, missing, or erroneous data. Data validation was most likely employed in the data collection process as there are no known outliers in the data. There are now 115,388 examples in this dataset with around 80% no citation issued and around 20% citation issued.

In the Police Use of Force data, there are around 100 observations with missing latitude and longitude, which are dropped given it is less than 1% of our entire dataset. Frequency tables were again created. There are around 300 observations with non-recorded race and 5 with non-recorded sex, which are dropped given it is around 1% of our

entire dataset. Erroneous precinct observations are dropped too. There are now 29,251 examples in this dataset with Bodily Force as the most common ForceType at around 75% of all examples.

4 Data Visualizations

Before modeling, we conduct preliminary descriptive analysis to understand our dataset and the relationships between variables better.

For the Police Stop data, we plot frequency plots for citation outcome by race, precinct, and problem. The following plots show that Asian, Black, East African, Latino, Other, and White have relative greater citations issued than not within their own race category. Overall, Black, Unknown, and White have the absolute highest citations issued. Precincts 1 and Precinct 2 have relative greater citations issued than not within their own precinct category. Overall, Precinct 2 and Precinct 4 have the absolute highest citations issued. Finally, Moving Violation has the highest relative and absolute, and Citizen/911 has the lowest relative and absolute citations issued. These could be important predictors in our later models.

For the Police Use of Force data, we plot frequency plots for force type outcome by race and precinct. The following plots show that Less Lethal and Less Lethal Projectile are most common Force Type outcomes for White, Unknown, Other/Mixed Race, and Native American. Comparatively, Gun Point Display and Firearm are most common for Black. Precincts also differ in Force Type outcomes, ranging from most common being Less Lethal or Less Lethal Projectile for Precinct 3 and Precinct 4 to Chemical Irritant and Maximal Restraint Technique for Precinct 1 and Precinct 2. These could be important predictors in our later models.

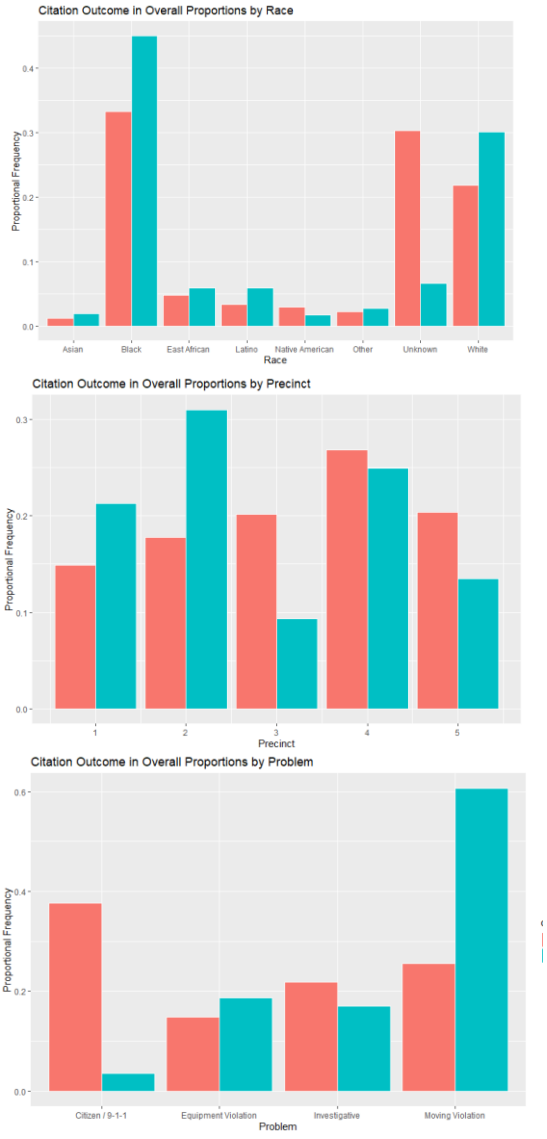
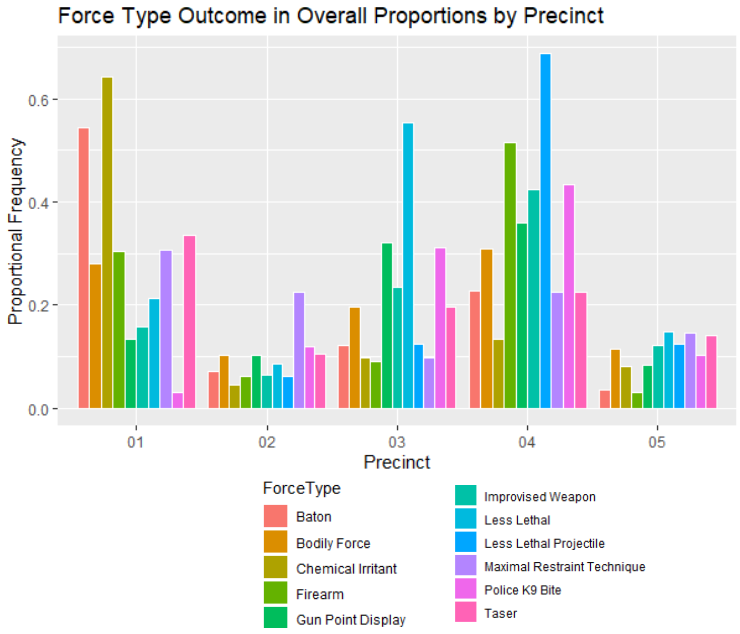


Fig 1.



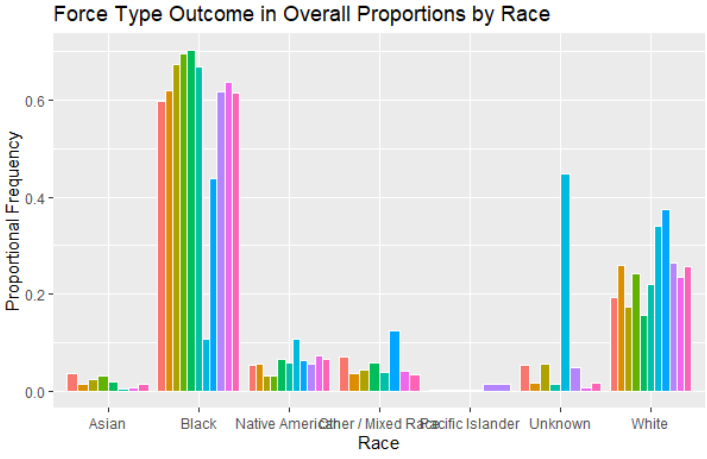


Fig 2.

5 Model Fitting

5.1 Police Stop

5.1.1 Logistic Loss

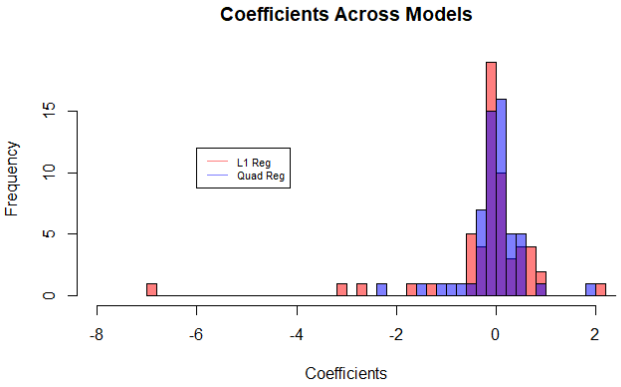


Fig 3.

Logistic loss with no regularization acts as our baseline model given this is a classification problem. We fit three different models of Logistic loss with no regularization, Logistic loss with L1 regularization, and Logistic loss with Quadratic regularization, and implement cross validation where 80% of the data is used as training across each of the 5 iterations, then the model is applied on the remaining 20% to obtain a validation error. To prevent overfitting, two regularizers are fit, where regularization parameters λ found using cross validation minimize misclassification. Figure 3 shows L1 regularization shrinks coefficients to 0 and has larger overall coefficient range, while Quadratic regularization has smaller range and less coefficients of 0. No regularization has the widest range, which is not pictured. Figure 4 shows no regularization has the greatest average misclassification error of .1521, followed by L1 regularization average error of .1517, then Quadratic regularization average error of .1516. Detailed analysis in Table 1 shows true positive, true negative, false positive, and false negative metrics averaged across the five validation folds.

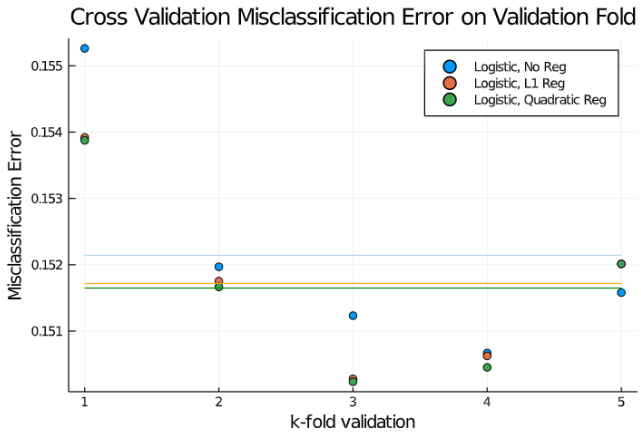


Fig 4.

		Actual		F1
Model		0	1	
Logistic, No Reg	0	19481.4	3417.8	.0459
	1	93.2	84.6	
Logistic, Quad Reg	0	19569.4	3496	.0036
	1	5.2	6.4	
Logistic, L1 Reg	0	19572	3497	.0030
	1	2.8	5.4	

Table 1.

Although Quadratic regularization has the smallest misclassification error, the model does poorly in classifying positives correctly because accuracy is largely attributed to the large number of true negatives in our data. To find a balance between classifying true positives and true negatives correctly, we turn to the F1 score.

$$F_1 = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

The F1 score for Logistic loss with no regularization is at .0459, then L1 regularization at .0036, and Quadratic regularization at .0030. Thus, Logistic loss with no regularization has the best performance, albeit still poor with such a low F1. Further modeling shows Hinge loss does similarly to Logistic loss in inadequately classifying true positives, and only achieves an average misclassification error of .1517 because it classifies true negatives correctly. Again, Hinge loss with no regularization has the best performance but with a severely low F1 score.

5.1.2 Classification Trees

We fit simple classification trees, random forest, and boosted models and implement 5-fold cross validation. For simple classification trees, the original models underfit severely by classifying all observations in the validation set as citationIssued = 0. Therefore, we specify a loss matrix that penalizes false negatives greater than false positives given our imbalanced dataset and model tendencies to classify observations as negative merely to minimize misclassification.

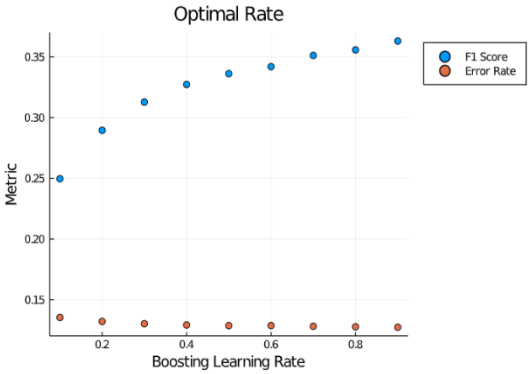


Fig 5.

An alternative tree method is creating ensemble models through random forest. This method creates many trees on the training set and combines the output together for a stronger classifier. Consequently, this technique reduces variance compared to a single classification tree, while maintaining low bias overall. For parameters, we use 500 classification trees, each with a maximum depth of 8 (square root of the 64 features). By limiting depth, we aim to prevent overfitted trees. By using a large number of classification trees, we aim to prevent underfitting.

In contrast, boosting increases the complexity of models that suffer from high bias, or weak learners that underfit the training data. To prevent overfitting the boosting algorithm, we introduce shrinkage via a small learning rate. Empirically, evidence shows using small learning rates less than .1 yields dramatic improvements in generalizability. We fit 5-fold cross validation boosting with learning rate ranging from .01 to .1, and find the optimal rate that maximizes F1 score for the validation fold equals .09 in Figure 5.

Figure 6 shows average misclassification error across the three methods, where the modified classification tree has an average of .2111, random forest an average of .1231, and boosting an average of .1273. Table 2 shows true positive, true negative, false positive, false negative, and F1 score metrics averaged across the five validation folds. We again see that the model with the smallest misclassification error is not necessarily the best performing model in F1; however, these models all perform notably better than the Logistic loss models in F1 where random forest and boosting have smaller misclassification errors, but the modified tree suffers a trade-off between maximizing F1 and minimizing misclassification error. Overall, random forest does the best balancing these two metrics.

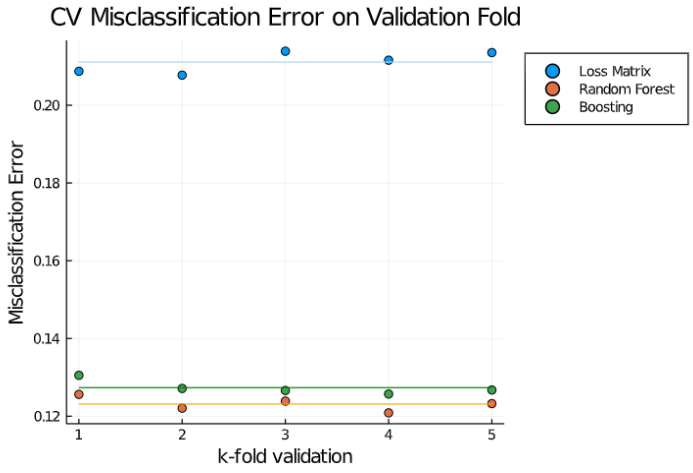


Fig 6.

		Actual		F1
Model		0	1	
Tree, Loss Matrix	0	16047.2	1344.2	.4697
	1	3527.4	2158.2	
Random Forest	0	19248.2	2515.2	.4099
	1	326.4	987.2	
Boosting	0	19300.6	2663.8	.3631
	1	274	837.6	

Table 2.

Figure 7 shows variable importance for classification tree with loss matrix, which refers to how much a given model uses each variable to make accurate predictions. The variables with the most predictive power include the problem situation, reason for the police stop, race, response date, and police precinct. Figure 8 shows variable importance for random forest, where latitude, longitude, response date, reason for the police stop, the problem situation, and race hold the most predictive power. Variable importance for boosting indicates similar results to random forest.

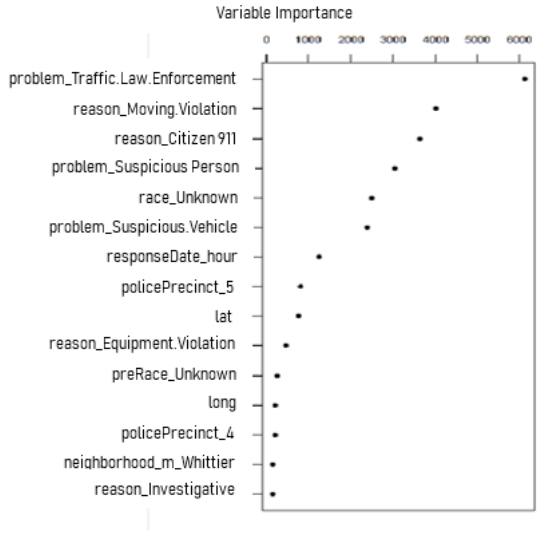


Fig 7.

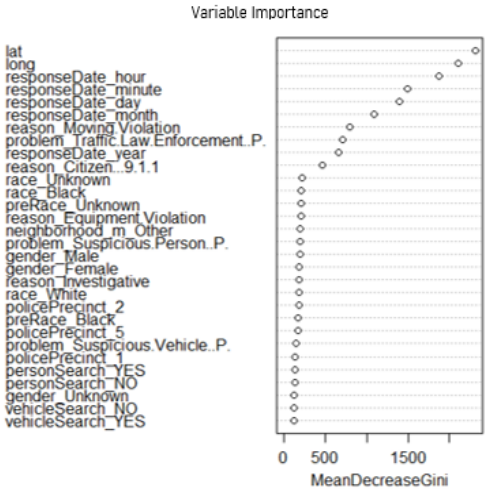


Fig 8.

5.2 Police Use of Force

5.2.1 Multinomial and One-vs-All Loss: First Attempt

Multinomial and One-vs-All losses with no regularization act as our first attempt for this multiclass classification problem. We fit three different models of Multinomial loss with no regularization, Multinomial loss with L1 regularization, and Multinomial loss with Quadratic regularization. We additionally fit three different models with One-vs-All loss in the same fashion as the models with Multinomial loss. Cross validation is again implemented where 80% of the data is used as training across each of the 5 iterations, then the model is applied on the remaining 20% to obtain a misclassification error. To prevent overfitting, two regularizers are fit, where regularization parameters λ found using cross validation minimize misclassification. When we originally ran these models, the predictions were skewed such that all observations in the validation set are classified as Bodily Force for its ForceType, giving us a misclassification error around .2733 for all six models. We found that similar to the Police Stop dataset, our Police Use of Force dataset is heavily imbalanced with around 75% ForceType labeled as Bodily Force out of 11 possible labels; consequently, models tend to classify observations as Bodily Force solely to minimize misclassification error.

5.2.2 Introducing SMOTE

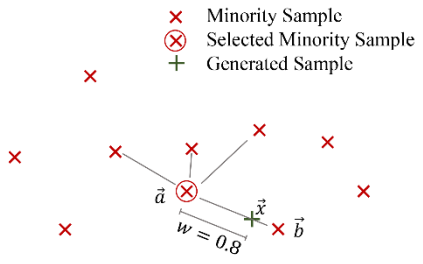


Fig 9.

To remedy this issue, we practice oversampling the minority classes and undersampling the majority class to compensate for the imbalance present in the data. Specifically, we follow a technique called SMOTE, or Synthetic Minority Oversampling Technique, which creates synthetic data by utilizing a k-nearest neighbor algorithm. It first selects a sample from the dataset, considers its k-nearest neighbors, and takes the vector between one of those k neighbors and the original data point. The vector is multiplied by a random number *w* between 0 and 1, then added to the original data point to create a new, synthetic data point[1]. Figure 9 illustrates the method where *k* equals 4, the selected neighbor is point *b*, and *w* equals 0.8.

Though a technique like stratified bootstrapping, where we sample with replacement from each sub-class independently with equal proportions, could be a possible solution to solve imbalanced data, it does not provide any additional information to the model. An improvement from simply duplicating data is to synthesize new data points. Thus, we use SMOTE with random undersampling of the majority class to create a more balanced dataset.

5.2.3 Multinomial and One-vs-All Loss: With SMOTE

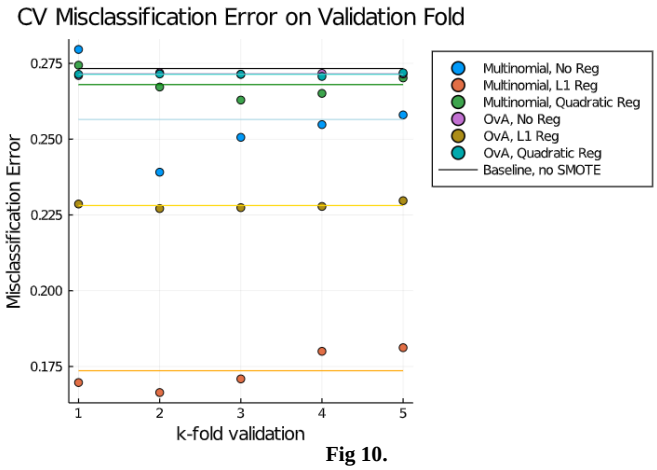


Fig 10.

After realizing we had a class imbalance problem, we created new datasets with our aforementioned SMOTE pipeline and reran all six models with these datasets. Figure 10 shows that for Multinomial loss, Quadratic regularization has the greatest average misclassification error of .2680, followed by no regularization average error of .2564, then L1 regularization average error of .1736. One-vs-All loss followed a similar pattern, with Quadratic regularization average error of .2715, no regularization average error of .2714, and L1 regularization average error of .2282. From this, we can see that L1 regularization performs the best out of the three regularizers, and overall the Multinomial model with L1 regularization performs the best with our data. Note that all 6 models achieved misclassification errors less than our first attempts without our SMOTE pipeline.

5.2.4 Black-Box vs. Glass-Box

Two types of modeling approaches are black-box and glass-box, where black-box models have observable input-output relationships but intermediate steps on how features are being combined to make predictions are unknown, while glass-box models have observable and understandable features and relationships. Glass-box models are typically more interpretable, but may suffer in accuracy compared to other deep learning algorithms. However, Explainable Boosting Machine combines interpretability through Generalized Additive Models with accuracy through modern machine learning techniques like bagging, gradient boosting, and automatic interaction detection to create performance comparable to techniques like random forests, and the ability to allow domain experts to correct for erroneous effects[2]. To test the performance between these two realms of modeling, we use SMOTE to create a more balanced dataset, and apply 5-fold cross validation to run random forest and EBM models. We also chose to not use the one hot encoding version of our data, as many of the features were categorical and we were more concerned with seeing the overall effect of these features as opposed to the differences within the categories.

Table 3 illustrates the comparison in validation fold errors. The average misclassification error for Random Forest is .2143, for EBM is .2113, and for Tree is .2308. Thus, we can see that not only is EBM more interpretable, but it actually has a higher accuracy as well. We also wanted to test EBM on the original dataset without SMOTE, to see how the model would handle imbalanced data. The subsequent misclassification error was .1998, which is even better than when we used SMOTE. This leads us to believe that the EBM model is well equipped to handle imbalanced data and overall does the best.

	k-fold					
Model	**1**	**2**	**3**	**4**	**5**	**Avg**
Random Forest	.2243	.1752	.2212	.2221	.2285	.2143
EBM	.2114	.2109	.2118	.2118	.2106	.2113
Tree	.2374	.2287	.2271	.2273	.2336	.2308

Table 3.

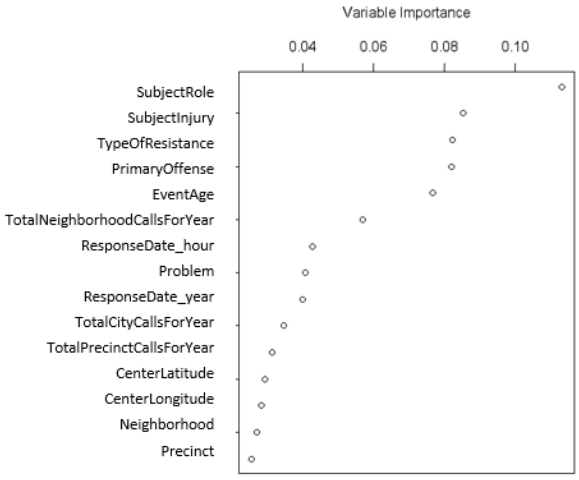


Fig 11.

Figure 11 shows variable importance for random forest in predicting ForceType. The most predictive variables include subject role, subject injury, type of resistance, and primary offense. Intuitively, these predictors make sense because they illustrate why certain types of police force would be used. For instance, a primary offense that is assault would more likely lead to Bodily Force from an officer than a primary offense of loitering. Figure 12 illustrates the comparative weights of the top 15 most influential features for EBM. The most predictive variables include subject injury, response date, primary offense, subject role, and situational problem. Again, this intuitively makes sense, as subject injury is a direct consequence of the force type. Additionally, other important features like primary offense and the problem at hand would inform the type of force that would be used.

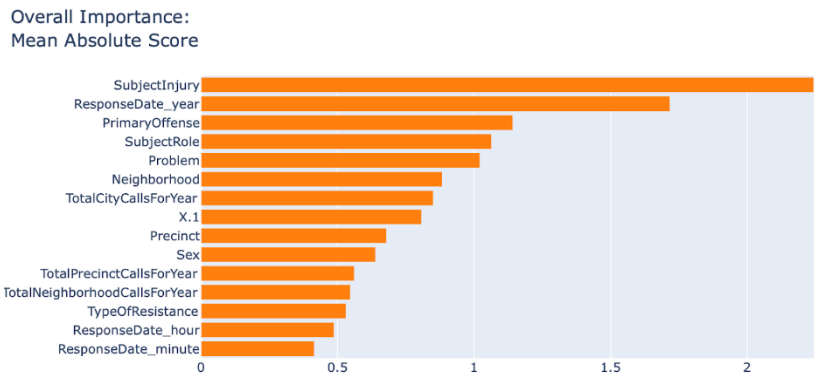


Fig 12.

6 Conclusion

6.1 Results Discussion

6.1.1 Model Review

For the Police Stop data, the random forest did the best in predicting if a citation was issued. While the classification tree with loss matrix had a better F1, ultimately random forest did better in balancing F1 score and minimizing classification error. For Police Use of Force data, the Multinomial loss with L1 regularization had the lowest misclassification error. However, the EBM model seems to do best overall, as it is both interpretable, has the second lowest misclassification error, and performs best on the unbalanced data. With a bit more tweaking to the graphs, we believe that EBM is the best model for predicting the type of police force that will be used.

As mentioned earlier, the Response year was the second-most influential feature in the Use of Force dataset. Figure 13 illustrates the score of each label throughout the different response years from the EBM glass-box model. We can clearly see that there is an inflection point around 2016-2017, specifically in labels 8 and 4, which correspond to Maximal Restraint Technique and Firearm respectively. Our original intention to gather Police data from Minneapolis was motivated by the recent killing of George Floyd in that city. During George Floyd's arrest, Derek Chauvin, a Minneapolis Policeman, knelt on Floyd's neck for nine and a half minutes, which is an example of maximal restraint. It is interesting to see that this model was able to learn that there was an increased usage of this technique on the part of the police, thus this unfortunate incident lies within the trend of our model. Additionally, the decreased usage of firearms follows the finding that in 2017, Minnesota had the

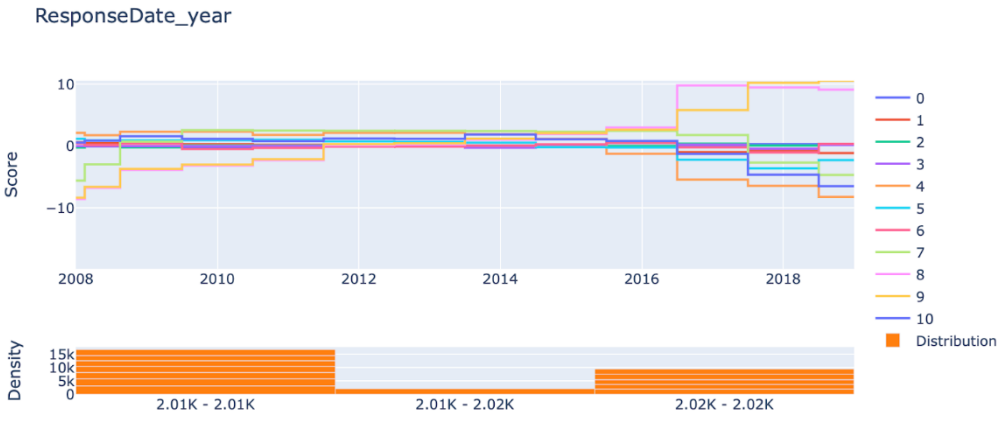


Fig 13.

eighth-lowest gun death rate in the country and exported crime guns at less than half the national rate[3]. These findings demonstrate the interpretability of EBMs, as with the other models we would not have been able to see such trends on a given feature.

6.1.2 Overfitting and Underfitting Analysis

Key concerns in modeling are the possibilities of underfitting or overfitting models on the training dataset. Overfitting occurs when a model incorporates too much noise from the data, and reduces generalizability onto the test data. A sign of overfitting is having low training error with high test error. Underfitting occurs when a model is not complex enough to accurately capture relationships between the predictors and target variable. A sign of underfitting is having both high training and test error. In contrast, generalizability occurs when a model has both low training and test error. We address these issues during model fitting by adding regularizers to loss functions, specifying a loss matrix for classification trees, and fitting random forest and boosted trees with optimal parameters. Table 4 shows the training and test error for one fold of 5-fold cross validation results on all models for Police Stop and Police Use of Force data. With comparable training and test errors, generalizability is adequately achieved.

Model	Train	Test		Model	Train	Test
Logistic, No Reg	.1513	.1552		**Multinomial, No Reg**	.2470	.2796
Logistic, Quad Reg	.1510	.1538		**Multinomial, Quad Reg**	.2663	.2744
Logistic, L1 Reg	.1511	.1539		**Multinomial, L1 Reg**	.1720	.1803
Tree, Loss Matrix	.2042	.2087		**OvA, No Reg**	.2621	.2715
Random Forest	.1093	.1256		**OvA, Quad Reg**	.2620	.2714
Boosting	.1248	.1305		**OvA, L1 Reg**	.2106	.2282
				Tree	.2292	.2374
				Random Forest	.2186	.2243
				EBM	.2107	.2133

Table 4.

6.2 Weapons of Math Destruction

Big data algorithms have increasingly been used in ways that reinforce pre-existing inequality. In our context, to prevent feedback loops that reinforce certain ForceTypes towards specific demographics or neighborhoods, these models serve to predict civilian-officer outcomes, but should not be used by officers or law enforcement to decide ForceType in specific future incidents. If our models were used in such a context, they would perpetuate pre-existing bias or unfairness present in past data that our models incorporate and become weapons of math destruction.

6.3 Fairness and Further Improvements

As we recently just learned about EBMs, we were excited to try to model our data with something that was both interpretable and accurate, especially given our trouble with imbalanced data. However, when running it, we found out that multiclass functionality in InterpretML is still experimental, thus further improvements in the multiclass classification functionality might improve the overall performance of the model. Given the features of both of our datasets, our models raise concerns of fairness or algorithmic bias. With more time, it would also be interesting to try to correct the model for bias or other possible errors with domain expertise that can be noticed by analyzing the individual effect graphs, as suggested by the glass-box methodology. In deciding the best model for civilian-officer outcomes, fairness or correcting bias could be an important criterion besides solely fitting correctly to past data.

Sources:

[1]

https://www.jair.org/index.php/jair/article/view/10302/24590?fbclid=IwAR0SJzmn1TUVE1Ys-csz9xini6lemjkP4ea1UCiLpwJu-G2bRMLOs9gC8Zs

[2]

https://arxiv.org/abs/1909.09223?fbclid=IwAR3S10rNy5-wMs3h_fmBkDH4uPEYYDNXBfWs2hdlx8GvqATV0pIdNNOYMZI

[3]

https://giffords.org/lawcenter/gun-laws/states/minnesota/