Checkpoint 4

Task #1 (Supervised): Could officer complaints be a 'canary in a coal mine' that could warn about future behavior? In other words, given an officer-filed allegation and the demographic, rank, and other information about the alleged officer, can we train a supervised model to predict whether that officer will receive a civilian complaint within 2 years?

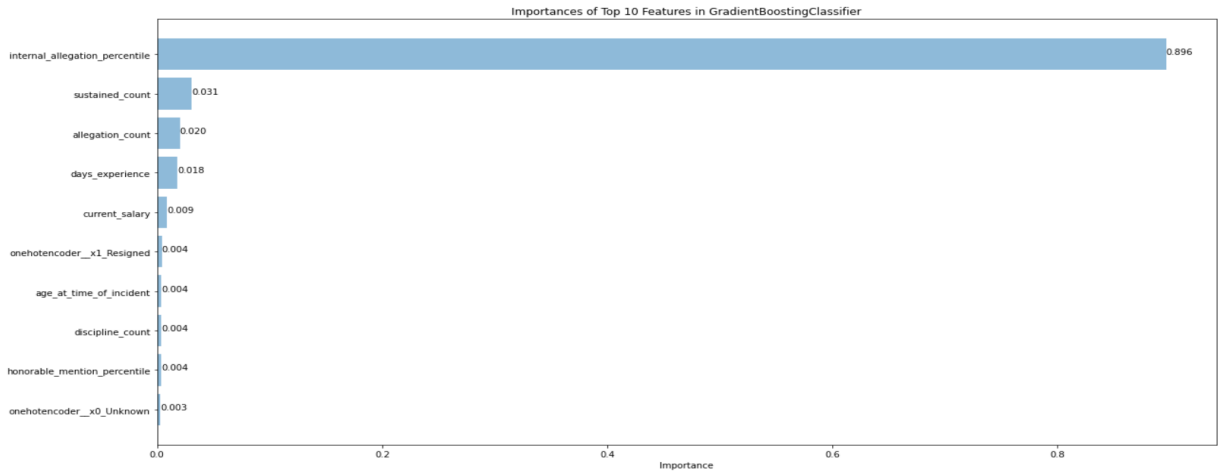
In this supervised ML task, we use officer-filed allegations and other information from the alleged officer to predict whether they would receive a civilian allegation within 2 years. The features we decided to use are basic information about the allegation, such as category, and information about the officer such as their salary, age at the time, and their allegation percentiles. We then train several different models to look for the best one using different models and cross-validation grid search. These models include logistic regression, decision tree, random forest classifier, gradient boosting classifier, and neural network. The best model is the gradient boosting model with the following parameter: {'max_depth': 1.0, 'min_samples_leaf': 0.01, 'min_samples_split': 0.01, 'n_estimators': 200}. Our model is able to reach approximately 81% accuracy, with an F1 score of 0.44. The model predicted the negative class (no civilian complaint in 2 years) with 82.8% accuracy and the positive class with 68.1% accuracy. While not an excellent score, this result is evidence that internal allegations are somewhat predictive of future civilian complaints.

Specifically, looking at the confusion matrix and its statistics for our model here, the precision for when the model is predicting not receiving future civilian complaints is exceptionally high (95%), suggesting that this model is excellent at identifying people who will not receive civilian complaints in the future. This even means that if we use this model to look for specific officers to pay attention to prevent future civilian complaints, we can be pretty confident that the officers that the models rule out are only 5% of the ones to receive civilian complaints even if we do not pay attention to them. The recall for class 1 classification is also decent (68%), suggesting that our model can identify potential trouble officers that will receive future civilian complaints, and such identification can definitely help signal for higher attention in order to prevent these future complaints. If this model were to be deployed for real-world use, we would want to achieve higher overall accuracy, but also raise accuracy on the positive class (current precision = 33%) so that there are more true positives, even if there are more false positives, since it may be better to flag officers incorrectly than to not flag at all.

		Truth data			
		Class 1	Class 2	Classification overall	User's accuracy (Precision)
Classifier results	Class 1	356	716	1072	33.209%
	Class 2	167	3450	3617	95.383%
	Truth overall	523	4166	4689	
	Producer's accuracy (Recall)	68.069%	82.813%		
Overall accuracy (OA):		81.169%			
Kappa[1]:		0.349			

Confusion Matrix (Class 1 = Have a civilian complaints in 2 years,
Class 2 = not have a civilian complaints in 2 years)

We also looked at the relative importance of each feature in the model. Most features had importance of zero and were not useful to the classifier, so we show the top 10 most important features in the chart below. It's possible that there are confounding factors that could explain the high importance of features like internal allegation percentile, but the presence of a variety of features on this list shows that the prediction is not entirely dependent on just a few features. One observation for the high importance of internal allegation percentile is that this means internal allegations are very important in these models to help predict future civilian complaints as this feature is directly related to the internal allegation. *(Note that the large size difference in importances is just an artifact of the calculation method, based on height in the decision tree, so relative importance is more important than absolute).*

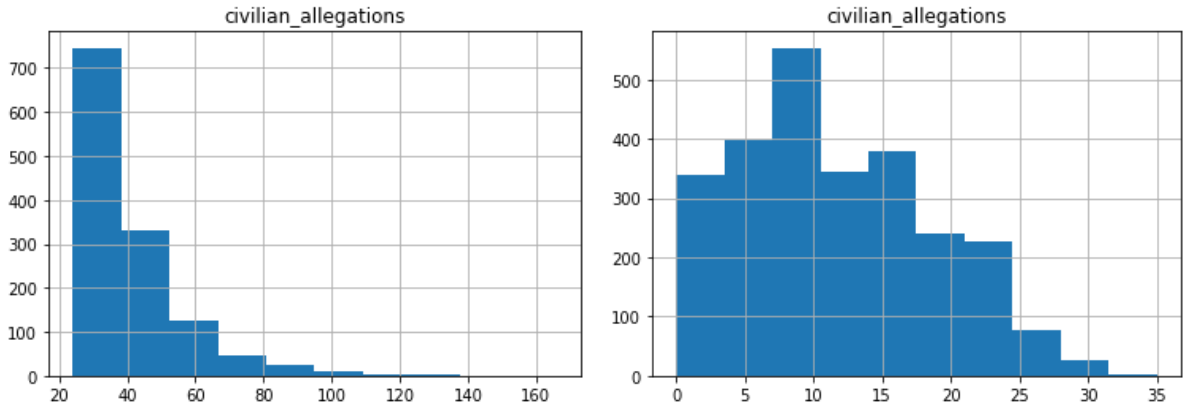


These results suggest that one can predict with decent accuracy whether an alleged officer will receive a civilian allegation in 2 years, based on information about the officer and the details of the allegation. Our hypothesis was that a given officer-filed allegation could be used as a “canary in a coal mine” to warn that an officer may receive civilian misconduct allegations in the future, and the moderate success of our model is evidence that supports this hypothesis. Even though the accuracy on the positive cases is only 66%, this still means that an officer allegation can point to future civilian allegations with some accuracy. For future work, since we train our model on both the officer’s information and the allegation details, we could separate the two to better understand whether the officer allegation is more predictive than traits of the officer themselves. This way, we might be able to predict future civilian complaints just based on details of the internal complaint, or just based on properties of an officer.

Task #2 (Unsupervised): We want to use unsupervised learning to cluster officers and find insights regarding the types of officers who have complaints filed against them by another officer, and to understand which traits are the most significant when determining what leads to a complaint.

We used K-means clustering to group officers who have internal complaints filed against them by other officers. The features that we decided to cluster on included: age, gender, years of service, race, number of civilian allegations, and number of awards acquired. Traditionally, K-means clustering is performed on numerical data but features such as gender and race are categorical. To address this, we use an algorithm called K-Modes which allows for the clustering of categorical variables. It works by defining clusters based on the number of matching categories between data points (https://github.com/nicodv/kmodes).

We tested different values of k for the clustering and using the elbow method we found that k=3 is the optimal number of clusters. This means that we can split up the officers into roughly three different groups. We then used a chi-squared test with the predicted labels to see which features were the most influential. The chi-squared scores that had the highest values indicated more important features within the clustering. We found that *age*, the *number of civilian allegations* that a police officer has accrued, and their *years of service* are the most telling when it comes to separating the clusters. For each cluster, we plotted a histogram of age, number of civilian allegations, and years of service to analyze how the groupings were made. Clearly, for each grouping, the features are very condensed. For example, in cluster 2, we see that the number of civilian allegations is no less than 20, but for cluster 3 we can see that there are very few civilian allegations above 25.



As for the features that did not receive a high score from chi-squared testing, we found the number of officer-filed allegations, gender, and race is not very significant. We can now be more confident that race and gender are not huge factors when trying to determine whether a police officer will obtain an officer-filed complaint, and focus more on the quantitative variables such as age and allegation count. Officer-filed allegations were the most surprising considering that the number of civilian allegations is important and basically the backbone of our project. We hypothesize this may be because civilian allegations are more common and thus are a greater tell of whether an officer will receive an officer-filed complaint behind the scenes. In other words, we should look further into whether officers have accumulated a lot of civilian allegations as this is a higher tell sign that they may perform misconduct worthy of an officer-filed allegation. This was alluded to in our interactive visualization. The converse may also be true where if an officer has been reported for misconduct among their peers, they are more likely to act similarly to the public.

The results from this machine learning model support our visualization from the last checkpoint, where we investigated relationships between certain officer traits and rates of internal complaints. In that visualization, there appears to be a clear trend for some categories like race and gender, but after normalizing, the trends disappear for those categories, indicating that they are not highly related to internal complaint rates. This is in contrast to the more significant features where even after normalizing, the trend remained consistent. One example is the high spike in 40-year-old officers both before and after normalizing. This is not necessarily an objective conclusive fact but was an analysis and trend that we recognized across our last two Checkpoints.