The Wicked Roadrunners
Kylie Chesner, Nimesh Tripathi, Aura Ulloa-Ordonez

Checkpoint 4 Findings: Machine Learning

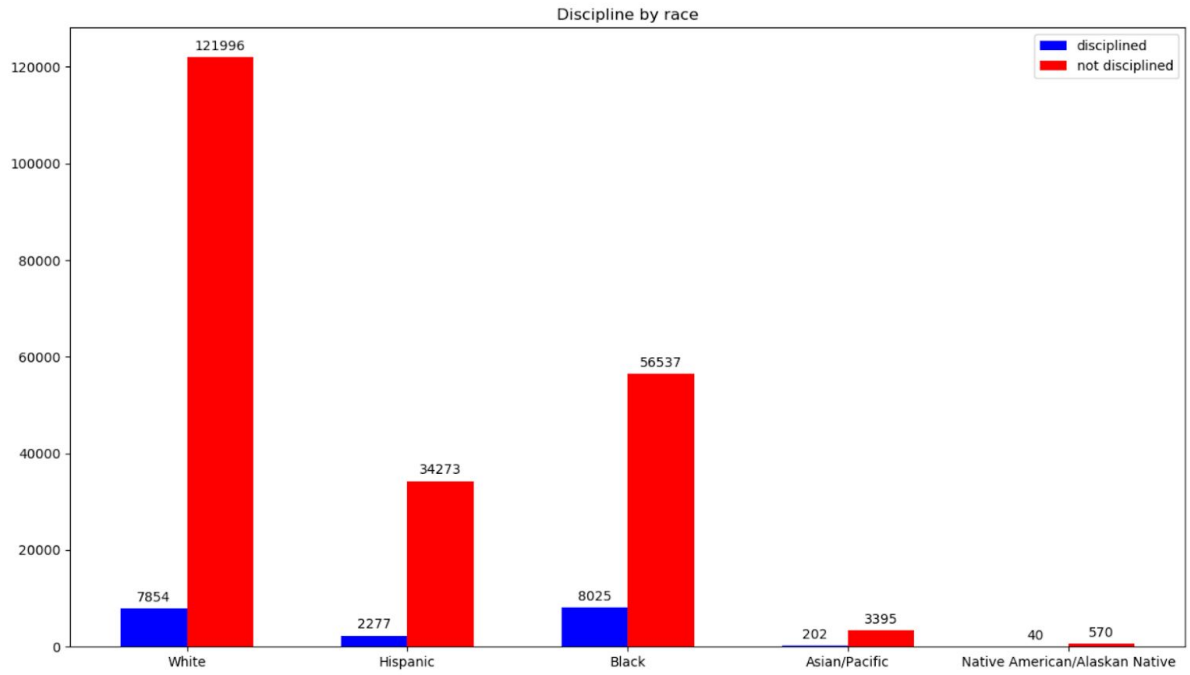
This checkpoint let us explore how certain officer or allegation characteristics relate to disciplinary action or settlement costs. This directly relates to our theme since it is investigating the relationship between an officer and acknowledgement of their wrong-doing, whether that be from discipline or settlements. The models and feature vectors can offer insight beyond our theme too, such as institutional biases present within the CPD. Additionally, the models could give civilians confidence to pursue an offending officer if they feel confident that officer will be punished and held accountable.

Question 1:

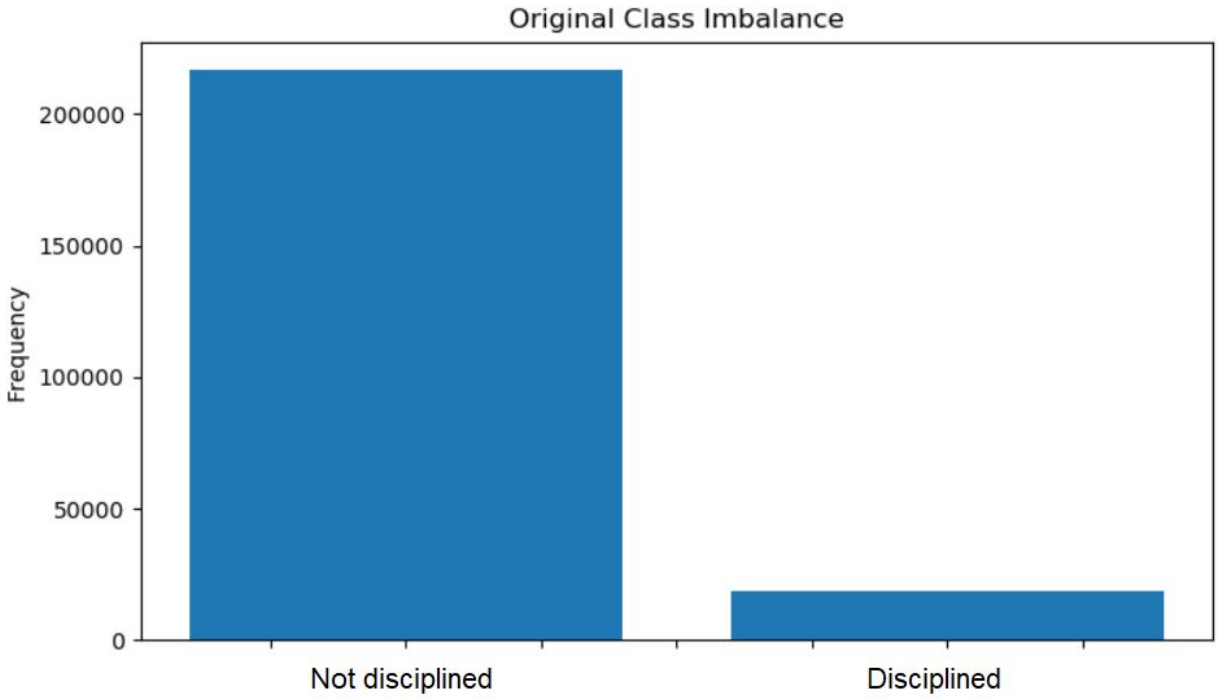
We wanted to investigate whether we could predict whether an allegation would lead to discipline based on officer and allegation data. If we can determine which combination of officer and allegation attributes lead to discipline, we can then further explore why some allegations are taken more seriously than others.Civilians may also be more likely to file a complaint if the model gives them confidence that the guilty officer(s) will be disciplined.

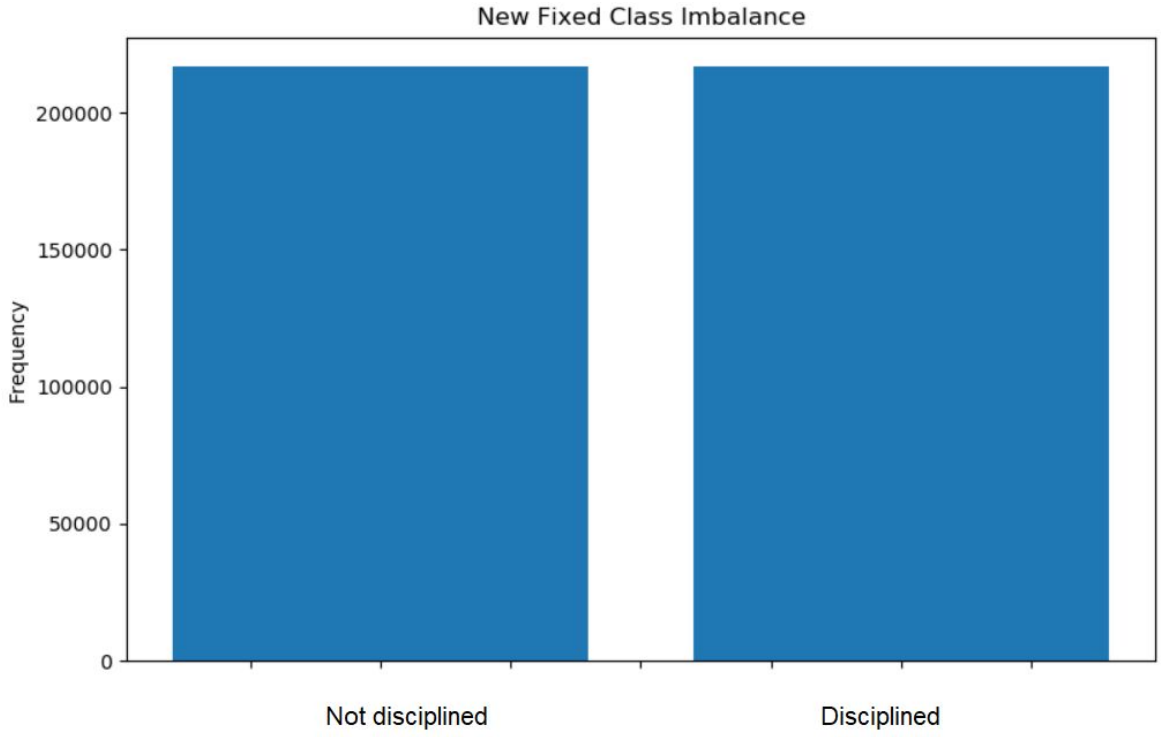
The fact that we can even predict what subset of allegations lead to discipline suggests some bias in the system. It suggests there are certain officer characteristics or certain allegation characteristics that let officers get away with misconduct.

Seeing which characteristics create the best model offers other unique insight into the CPD. We found that the features race, rank, complaint percentile, allegation count, and misconduct category gave the best performance for predicting discipline. In our checkpoint 2 findings, we saw that not all officer ranks receive discipline proportionally; thus, it isn't surprising that using *rank* as a feature strengthens our model. Allegation characteristics make sense for the predictions (e.g. excessive force should be disciplined more harshly than an operation violation). On the other hand, a relationship between officer race and disciplinary action implies racial biases within the Chicago police institutions. In the figure below, we see black officers are disciplined more than white officers, meaning white officers are able to get away with more misconduct. Officer gender didn't impact the model performance much, which is why it was omitted. We assumed this is because most cops are male and male officers tend to commit more violent misconduct.



Before we could create our model, we needed to perform some data cleaning and processing. First, there is a large class imbalance, as a large percentage of all allegations do not lead to discipline. To fix this, we oversampled the minority class.





After resampling, we needed to preprocess our data into a usable form for the Sklearn models. We encoded all categorical input features as numbers using Sklearn's preprocessing.LabelEncoder() method.

We tested multiple classifiers to find the best option for our data. We tried MLP, Naive Bayes, k nearest neighbor, and AdaBoost, but ultimately found that Random Forests give us the best results. To determine the optimal number of trees in the forest, we kept increasing the hyperparameter until performance plateaued; this value was 15.

We evaluated our model two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy. The image below shows these results.

	precision	recall	f1-score	support
False	0.961	0.811	0.879	43302
True	0.837	0.967	0.897	43407
accuracy			0.889	86709
macro avg	0.899	0.889	0.888	86709
weighted avg	0.899	0.889	0.888	86709

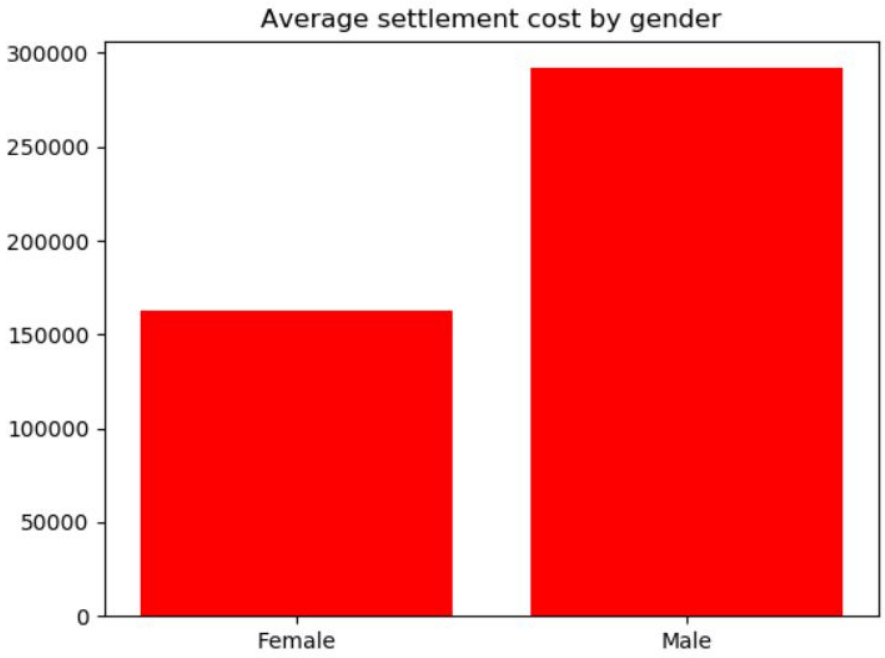
We also used 5-fold cross validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.88695522 0.88499464 0.89054067 0.89162476 0.88187941].

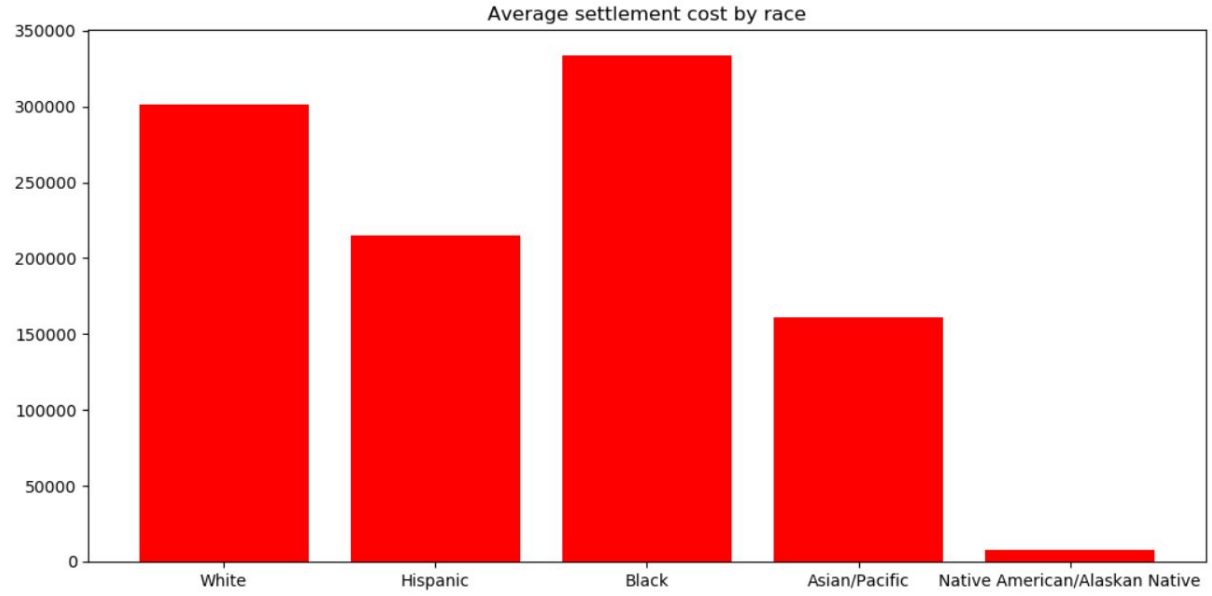
Thus, we can confidently say our model was a success and can predict with 89% accuracy whether an officer will be disciplined for an allegation. However, we were surprised with the impressive performance of our model. Since very few officers are held accountable for misconduct, we expected the data to be more random and noisy. Instead, we see that there is a pattern to officer accountability. In the future, it would be interesting to investigate how well the input features can predict the severity of the discipline (e.g. verbal warning vs suspension).

Question 2:

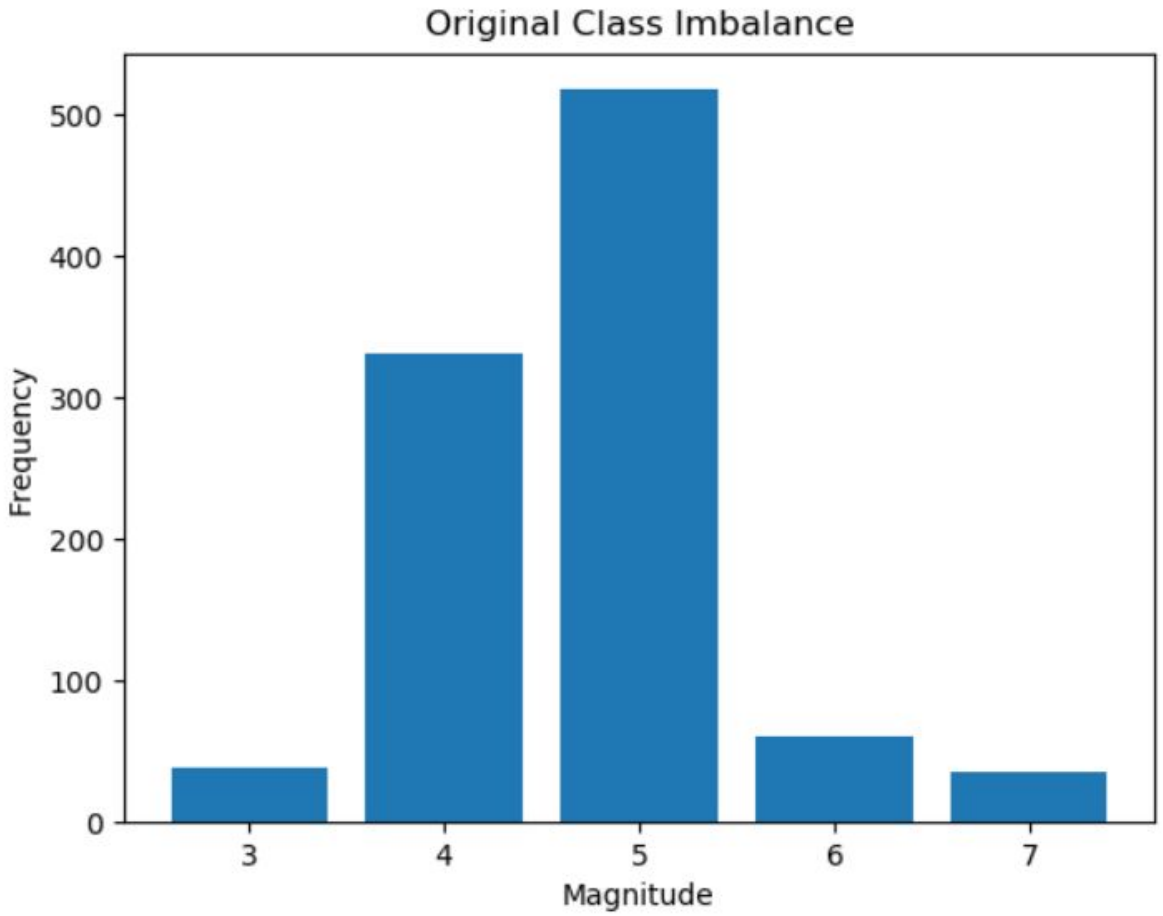
We wanted to see if we could predict the settlement amount of an allegation lawsuit based on officer and allegation characteristics. Since a settlement amount difference of a few dollars is insignificant, we tried to predict the magnitude of the settlement amount. For example, $10 has a magnitude of 1, $100 has a magnitude of 2, $1000 has a magnitude of 3, etc. The reasoning behind this is if we can see which officer or allegation characteristics cost the CPD the most money, we can identify problematic areas in the CPD and lawsuit process. This model would also be valuable to civilians, as they could assess whether it is worth it to try to settle for money in court. If the settlement amount is predicted to be low, then the process might not be worth it with all the time and legal fees involved.

The input features we used for the model were: interactions (e.g. traffic stop, home invasion), outcomes (e.g. killed by officer, drug charges), misconduct category, gender, and race. In our checkpoint 2 findings, we saw that settlement costs varied highly by misconduct category. Thus, it makes sense that this attribute enhances our model. We can see from the plot below that male officers have around twice the average settlement cost as females. We suspect this is because most cops are male and male officers tend to commit more violent misconduct. We also see from the figure below that black officers have the highest average settlement cost, shortly followed by white officers. Other racial groups have significantly lower average settlement costs. These plots helped us identify the important officer characteristics to use in the model. The other features are related to the allegation and are more intuitive. For example, we would expect being killed by an officer to result in higher settlement than resisting arrest.





Before we could create our model, we had to perform some data preprocessing. We first had to transform the settlements to their magnitude using the Pandas dataframe transform method. Then, similar to Question 1, we had to resample to fix class imbalances (see figure below). A magnitude of 5 was the majority class, so we oversampled the other classes to match.



Similar to Question 1, we also tested multiple classifiers to find the best option for our data. We tried MLP, Naive Bayes, k nearest neighbor, and AdaBoost, but ultimately found that Random Forests give us the best results. To determine the optimal number of trees in the forest, we kept increasing the hyperparameter until performance plateaued; this value was 10. We also chose to use Random Forests because they are less influenced by outliers, ensemble models are more resistant to overfitting, and they still work with correlation among input features.

We again evaluated our model two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy. The image below shows these results.

	precision	recall	f1-score	support
3.0	0.940	1.000	0.969	109
4.0	0.925	0.916	0.920	107
5.0	0.989	0.860	0.920	107
6.0	0.925	1.000	0.961	98
7.0	1.000	1.000	1.000	97
accuracy			0.954	518
macro avg	0.956	0.955	0.954	518
weighted avg	0.955	0.954	0.953	518

Second, we used 5-fold cross validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.80694981 0.88223938 0.90926641 0.85521236 0.84555985].

One thing we noticed was that for Question 1's model, the classification report and cross validation numbers were the same (both around 89%), but for Question 2, the classification report shows a slightly higher performance (95% vs 85%). This implies a small amount of overfitting. However, since the cross validation numbers are still quite high (between 80 and 90%), we still would say our model is successful in its ability to predict settlement money magnitudes.

While both models show good performance evaluation, these models are only as good as the data used to train and test them. Because we had to resample due to class imbalances, our data might not be truly realistic. However, these models provide good starting points for further explaining trends in the CPD relating to discipline and settlement.