

Stage IV: Predicting NYPD Misconduct Case Outcomes and Penalties

Helen Wang

Literature Review

In 2020 alone, 53.8 million U.S. residents reported having contact with law enforcement (Tapp & Davis, 2020). While the majority (88%) of respondents reported having satisfactory interactions with police, about 2% or 1 million respondents reported having negative experiences, characterized by the threat and/or use of force as well as other forms of misconduct (Tapp & Davis, 2020). Police misconduct, defined by the New York Police Department (NYPD) Civilian Complaints Review Board, is any action or charge that may result in an officer being subjected to an administrative trial process (Cubitt et al., 2022). Conduct that falls under this category includes, but is not limited to, improper use of force, abuse of authority, discourteous behavior, and offensive language (Cubitt et al., 2022). In recent years, police misconduct has become a prominent issue in the eyes of the American public, especially given racial/ethnic disparities in experiences, with numerous high-profile cases resulting in the death of many lives. Releases of police misconduct records in recent years have allowed researchers to better characterize and understand the escalating problem at hand.

Research has shown that there are several correlates of police misconduct. In general, male officers are more likely to get complaints than their female counterparts (Harris & Worden, 2014). Younger, less experienced officers are also at higher risk along with officers with military experience (Harris & Worden, 2014). Officers who are more productive (i.e. with more arrests and stops) and who have more civilian interaction are also more likely to accrue greater complaints of police misconduct (Harris & Worden, 2014; Rozema & Schanzenbach, 2019). Some reports have also suggested that officers from minority populations may also be more likely to commit misconduct, though other studies have suggested that this may be a result of differential task or geographic assignment to high-crime precincts (Cubitt et al., 2022; Harris & Worden, 2014).

Police misconduct records only represent a small portion of actual misconduct cases; it is estimated that approximately only a third of complaints end up being actually filed (Harris & Worden, 2014). Of those filed, legal and institutional barriers make it difficult for complaints to result in disciplinary measures. Reports have shown that only an eighth of civilian-initiated complaints are substantiated, with most complaints often being declared as either exonerated (act verified but found to be proper), unfounded, or not substantiated (insufficient evidence) (Harris & Worden, 2014). If substantiated, about 1 of 24 cases result

in sanctions for officers involved, with most sanctions often not commensurate with the misconduct that occurred (Harris & Worden, 2014).

Past work looking into police misconduct at various police departments across the nation has identified significant patterns. Rozema & Schanzenbach (2019), while examining police misconduct in Chicago, found that officers with moderate numbers of misconduct allegations were at no greater risk of committing serious misconduct than officers who had no misconduct allegations at all. Instead, they found that officers who were in the top 1% quartile of misconduct allegations were more likely to commit serious misconduct, generating almost 5 times the number of payouts and 4 times the total damage payments. When complaints against officers are substantiated, almost 36% of offending officers accrue another substantiated complaint at some point during their career (Harris & Worden, 2014). The risk of obtaining another complaint was found to be higher in the first months succeeding the first complaint but dropped significantly afterward. Notably, officers who are sanctioned for the complaint are not only more likely to engage in misconduct but do so more rapidly (Harris & Worden, 2014). Overall, findings suggest that a small subset of repeat offenders are responsible for a large portion of police misconduct reported, encompassing over \$1.5 billion in lawsuit settlements across the nation (Cubitt et al., 2022).

Cubitt et al. (2022) also noted that differences in case outcomes depend on officer and complainant characteristics. Female officers, for example, while less likely to accrue misconduct allegations in general than their male counterparts, were more likely to be sanctioned with remedial management action. Black and Hispanic civilians who submitted a complaint of police misconduct were 4.7 and 1.6 more likely to receive a not substantiated ruling compared to White citizens (Headley et al., 2020). One study also found that racial mismatches between officer and complainant were linked to differing case outcomes (Wright II, 2020). The study found that Black complainants were more likely to receive a substantiated ruling when misconduct was alleged against a white officer. On the other hand, white complainants were less likely to receive a substantiated ruling when alleging misconduct against a black officer. These results differed across city departments, however, suggesting geographic differences.

Research Question(s)

The issue of police misconduct is important, especially given rising cases of police violence, police brutality, and fatalities as a result of such misconduct. Such issues have resulted in the loss of multiple lives, increased racial tensions across the US, and fractured public trust in law enforcement. The literature has explored various factors linked to police

misconduct and has similarly evaluated predictors of not only future misconduct but also misconduct case outcomes.

Few studies, however, have examined the impact of media coverage on police misconduct. Those who do have primarily limited their analysis to the impact of media coverage on public perceptions of police misconduct (Chermak et al., 2006; Dowler & Zawilski, 2007). No article to date has assessed the role that the media can play in influencing misconduct case outcomes, despite the potential pressure that media can have on ensuring that sufficient sanctions are levied.

To address this gap, in this study, we will utilize a dataset of police misconduct cases drawn from the New York City Police Department (NYPD) to do the following:

1. Develop supervised models to predict case and penalty outcomes for police misconduct cases.
2. Assess the impact that media coverage/visibility of a given police misconduct incident has on its case outcome, specifically its complaint disposition, as well as the type of resulting penalty.
3. Identify and verify the factors most predictive of police misconduct case outcomes as well as penalty outcomes.

Data

To answer these questions, I will utilize police misconduct cases from the NYPD. Established in 1845, NYPD is one of the oldest and largest police departments in the nation, encompassing over 36,000 officers and 19,000 civilian employees (*About NYPD*, n.d.). Across its 78 precincts, the department serves over 8.5 million different individuals. The Civilian Complaint Review Board (CCRB), which separated from NYPD in 2000, has compiled a database of over 395,000 police misconduct cases from 2000 to 2025 (*Civilian Complaint Review Board (CCRB) Database*, n.d.). The database contains 4 datasets:

1. Allegations Against Police Officers: a list of all closed allegations made against NYPD officers, including information about the complainant, the officer, allegation, and resulting disposition (*Civilian Complaint Review Board: Allegations Against Police Officers | NYC Open Data*, n.d.)
2. Complaints Against Police Officers: a list containing information such as dates, locations, and circumstances surrounding the allegation (*Civilian Complaint Review Board: Complaints Against Police Officers | NYC Open Data*, n.d.)
3. Police Officers: a list of all NYPD officers and the number of total and substantiated complaints on their record (*Civilian Complaint Review Board: Complaints Against Police Officers | NYC Open Data*, n.d.)

4. Penalties: a list containing case and trial penalty information (*Civilian Complaint Review Board: Penalties | NYC Open Data*, n.d.)

To get media coverage information, I will use a subset of data available from the Mapping Police Violence Project, which used google alerts to get news articles on police violence events to construct their dataset (*Mapping Police Violence*, n.d.-a). Currently, the Mapping Police Violence Project has records of 48 separate police violence incidents, with associated news article links, that occurred between 2013 and 2024 (*Mapping Police Violence*, n.d.-b). Given the sample size, I will also utilize the New York Times API to scrape for relevant news articles on the officers and their associated misconduct cases.

Preprocessing and Descriptives

Given the sheer number of datasets I am utilizing for this study, merging datasets effectively is a critical component. I first merged the Allegations Against Police Officers and Complaints Against Police Officers datasets together by inner join, as there are multiple allegations forming single complaints (i.e. allegation ID is clustered by complaint ID). I then merged the resulting dataset to Penalties by left join, since Penalties are only given to complaints that have been ruled as “Substantiated” and I want to keep non-substantiated case outcomes. Finally, I merged this result with Police Officers by inner join. Police Officers contain all officers on the NYPD roster, regardless if they have a complaint against them. For this study, as I am interested in exploring case outcomes and resulting penalties, I used inner join to drop all officers who have not been subject to a complaint/allegation. This left me with a resulting dataset of 235,939 instances.

As this dataset lacked any variables for media coverage, I utilized the Mapping Police Violence Project dataset. After filtering for instances from the NYPD and for cases where the officers involved were known, I merged by officer name and year of incident using left join. Because the Mapping Police Violence Project dataset runs from 2013-2024 and the NYPD CCRB runs from 2000 to 2025, I elected to filter for instances that occurred after 2013. I also filtered for cases that occurred before 2020 to avoid potential endogeneity from COVID as well as the Black Lives Matter Movement (which was sparked in part by several police brutality and murder incidents that occurred around the time, which would likely skew media coverage).

For our main target variables, I am focusing on case outcome (CCRB Complaint Disposition) and resulting penalty (NYPD Officer Penalty). I utilized the CCRB Complaint Disposition rather than the individual allegation dispositions as (1) only Substantiated Complaints rather than Substantiated Allegations result in a penalty, and (2) multiple allegations are clustered under individual complaints. I also used the NYPD Officer Penalty

instead of other penalty variables as they are typically recommended penalties and NYPD has final authority on which penalties are actually implemented.

Upon examination of the CCRB Complaint Disposition variable, in addition to expected disposition types (Substantiated, Unsubstantiated, Exonerated, Unfounded) there were also miscellaneous types such as “Complainant Uncooperative”, “Complaint Withdrawn”, “Subject Resigned”, etc. For the miscellaneous types, as they effectively meant the disposition did not happen, I recoded them as NaN and dropped those rows, resulting in a dataset of 36,666.

The NYPD Officer Penalty variable was also highly variable and in string format. In terms of penalties, some officers received multiple types and for different durations. For analysis, I engineered a variable capturing the total number of different types of penalties each officer accrued. Simultaneously, I recoded Penalties into a categorical variable, with multiple penalties recoded to the most severe penalty applied.

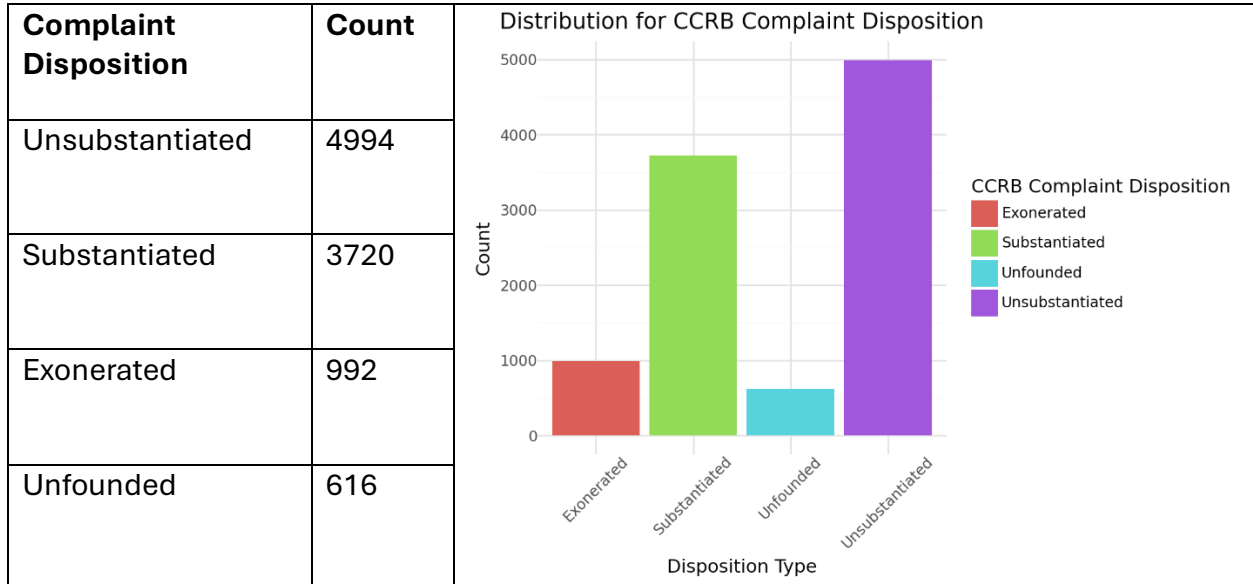
Using only the Mapping Police Violence Project, there are only roughly 33 instances where there was media coverage over the period of interest. Due to the scarcity of instances that fell under this criteria, I also incorporated data from the NYT Article Search API. Given rate limits, I elected to take a random subset of 4000 rows from the merged dataframe and extracted unique officer names and years. Using the officer name and the search terms “AND (police OR officer OR NYPD) AND (misconduct OR force OR brutality OR violence)” while filtering for articles within a year of the misconduct instance, I scraped for any hits. I then inner joined the results to the merged dataframe, giving me a resulting dataset of 10322, with the numbers reflecting the fact that multiple allegations are being clustered under singular complaints. While smaller than the full dataset, scraping hits for all misconduct cases would have been impossible—as such, I elected to look only at 4000 randomly selected cases.

To account for temporality with the data, I also engineered three variables that captured the occurrence of the following in the last 2 years: (1) a previous complaint/allegation, (2) a case where the outcome was “Substantiated”, and (3) a case that had resulted in an actual penalty. Previous literature had noted that misconduct is more likely within the first couple of months following a previous case—as such, I limit our surveillance to within the last 2 years.

Furthermore, I also utilized z-score normalization for numerical data variables and (depending on data distribution), reset intervals for categorical variables prior to dummy coding. After cleaning, we had minimal missing data: Incident Hour (0.4%), Precident of Incident Occurrence (1.7%), and Incident Month (0.1%). Data was imputed using MICE.

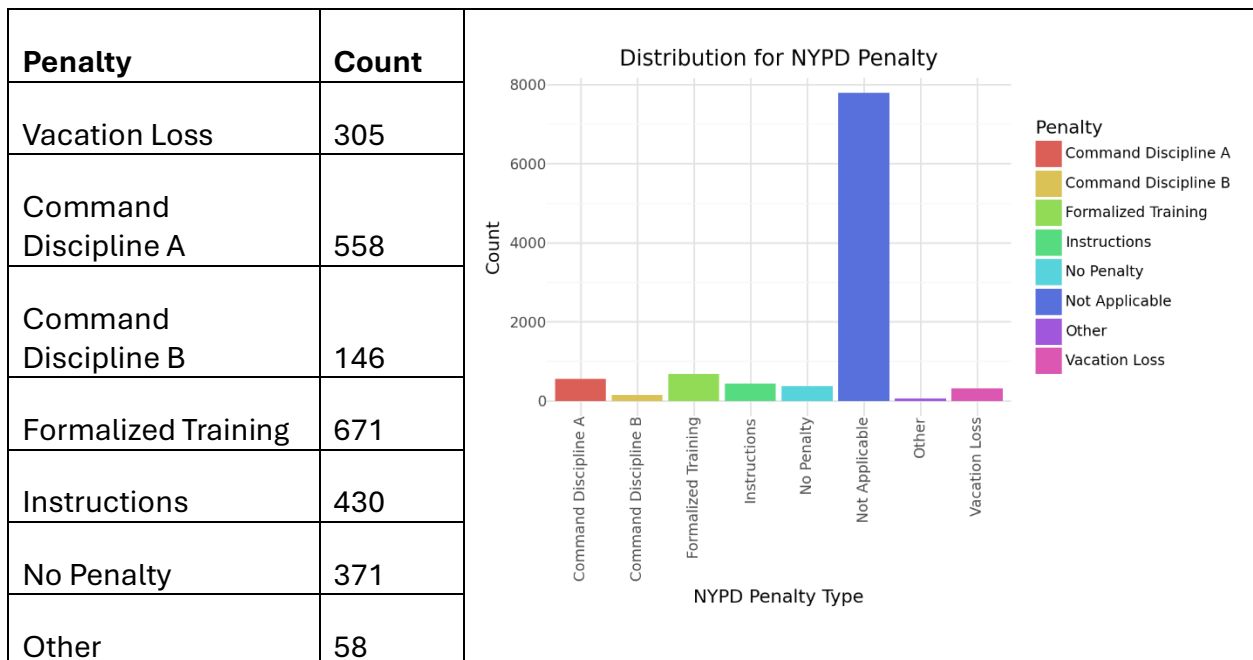
With the reduced dataset, I examined the distributions for the target variables below. In all cases, I observed significant imbalance.

Table 1. CCRB Complaint Disposition



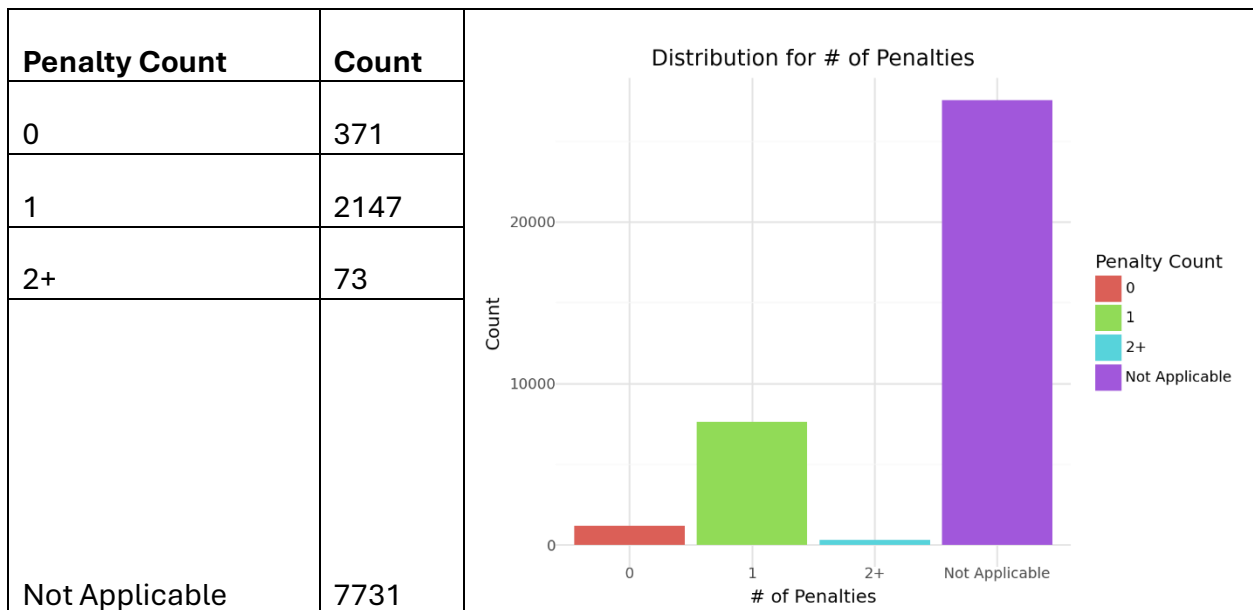
In the case for NYPD penalty, since penalties are only ever allotted if a case is “Substantiated,” there are cases where they are “Not Applicable.” When running models for these target variables, I limit the dataset to only cases where penalties are applicable.

Table 2. NYPD Officer Penalty



Not Applicable	7783	
----------------	------	--

Table 3. NYPD Penalty Count



Given the imbalance present in our target variables, I utilized synthetic minority oversampling technique (SMOTE) on training sets prior to running models..

Methodology

My main feature of interest is a dummy variable indicating whether there was media coverage or not. Additional features that will be assessed include but are not limited to: Officer Days on Force at the Incident, Precinct of Incident Occurrence, Officer's Total Complaints, Officer Rank at Incident, Type of Misconduct, Victim demographics, Officer demographics, Location of Incident, etc.

Our main target variables in this case are the categorical variables: CCRB Complaint Disposition and NYPD Officer Penalty (captured via two categorical variables, Penalty Recoded and Penalty Count). For the first target variable, models need to be classifiers that are optimal for predicting multiclass variables and ideally are not sensitive to imbalance (given that is our biggest problem with our dataset). To account for imbalance in model performance, I will utilize F1-scores (weighted and macro) as well as balanced accuracy to assess model performance. Simultaneously, I will use confusion matrixes to assess model performance for each class.

For our second target, NYPD Officer Penalty, given that officers can receive multiple penalties (i.e. loss of vacation and formalized training), we need models to (1) predict the type of penalty and (2) predict the number of penalties. Models would need to be optimized

for multiclass classification to capture both aspects and similarly need to be less sensitive to imbalance. As such, I used similar metrics as mentioned for the first target.

In all cases, I will avoid using models such as Decision Tree, due to its sensitivity to imbalanced data. Simultaneously, I will also avoid models such as Naïve Bayes, given that some of the model assumptions (feature independence) may be violated. Instead, I will utilize models such as logistic regression, support vector machine, k-Nearest Neighbors, Random Forest, and Gradient Boost. After splitting data into a 60-20-20 split for training, validation, and testing, I will run initial models. Depending on model performance, I will tune hyperparameters for a select subset of high-performing models on the validation set using randomized search (to reduce computational expense) and cross-validation.

Results

I ran 5 different initial models to predict case outcome. Only two of these models (both ensemble method models) showed performance that was better than chance when utilized against the validation set. Examination of the confusion matrixes suggests that models are doing well in predicting "Unsubstantiated" (Class 3) cases but struggle with the other categories. Random Forest and XG Boost appear to be moderately successful with predicting "Substantiated" (Class 1) in addition to "Unsubstantiated" (Class 3) cases but also struggle with "Exonerated" and "Unfounded" cases (specifically for recall rather than precision). The two classes, "Exonerated" and "Unfounded," were the primary source of imbalance in our dataset, being the minority classes. Despite efforts to correct this via SMOTE, it appears that it still affected model performance.

Using randomized search and cross-validation, I attempted to identify the optimal hyperparameters for Random Forest and Gradient Boost. After tuning, best parameters resulted in an F1 Score (weighted) of 0.886 for our Random Forest model and an F1 Score (weighted) of 0.891 for our XG Boost model. Though differences were minimal, Gradient Boost appeared to be performing the best out of the models tried. When evaluated on the test set, I obtained a F1 Score (weighted) of 0.85. Evaluating feature importance suggested that the number of total substantiated complaints, body camera evidence, and type of abuse of authority played the biggest role. Notably, media coverage was not within the top 20 features in terms of importance, suggesting minimal influence on case outcome.

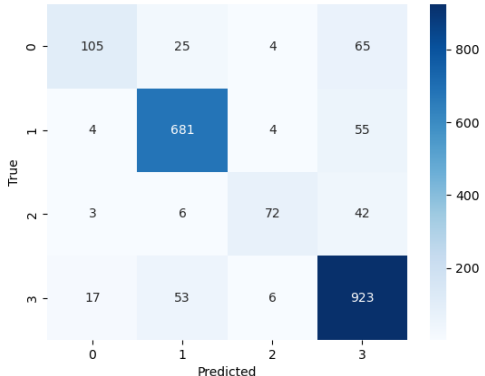
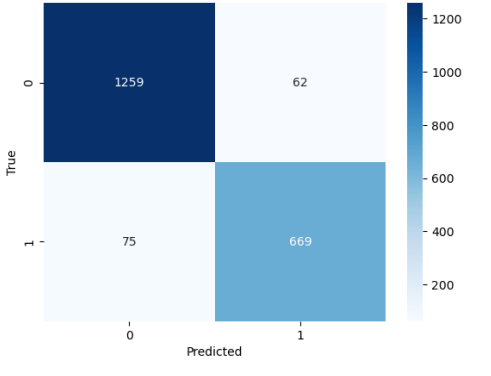
Given the models limitations in estimated "Exonerated" and "Unfounded" cases, I elected to assess if dichotomizing the target variable would improve model performance. Specifically, I elected to recode the target variable as "Substantiated" (1) or "Not Substantiated" (0), as only the former would result in a penalty. I ran the same initial models and found that while performance did improve marginally for all models, ensemble

method models still massively outperformed the rest. After hypertuning parameters, differences between Random Forest and Gradient Boost model performance were minimal at an F1 Score (Weighted) of 0.924 and 0.927 respectively.

Gradient Boost performed slightly better however so I elected to utilize that to compute performance on the test set. Examination of the final confusion matrix suggests that the model is slightly better at detecting True Negatives versus True Positives. Simultaneously, it appears more prone to False Negatives than False Positives. The overall performance of the final model was acceptable however, with a F1 Score (Weighted) of 0.933. Examination of feature importance showed that similarly, the number of total substantiated complaints and body camera evidence played a critical role. For the binary model however, officer rank was also important. Media Coverage was also not in the top 20 most importance features.

Overall, a binary model appears to have slightly better performance in terms of predicting case outcomes but both models had acceptable performance. See specifics in the table below:

Table 4. Multiclass and Binomial Models for Complaint Case Outcomes

Model	Multiclass	Binomial																																		
F1 Score (Weighted)	0.846	0.933																																		
Confusion Matrix	<div><p>Confusion Matrix</p><p>A 4x4 confusion matrix for a multiclass model. The x-axis is 'Predicted' (0, 1, 2, 3) and the y-axis is 'True' (0, 1, 2, 3). The diagonal elements are 105, 681, 72, and 923. A color bar on the right indicates counts from 0 to 800.</p><table><thead><tr><th></th><th>0</th><th>1</th><th>2</th><th>3</th></tr></thead><tbody><tr><th>0</th><td>105</td><td>25</td><td>4</td><td>65</td></tr><tr><th>1</th><td>4</td><td>681</td><td>4</td><td>55</td></tr><tr><th>2</th><td>3</td><td>6</td><td>72</td><td>42</td></tr><tr><th>3</th><td>17</td><td>53</td><td>6</td><td>923</td></tr></tbody></table></div>		0	1	2	3	0	105	25	4	65	1	4	681	4	55	2	3	6	72	42	3	17	53	6	923	<div><p>Confusion Matrix</p><p>A 2x2 confusion matrix for a binomial model. The x-axis is 'Predicted' (0, 1) and the y-axis is 'True' (0, 1). The diagonal elements are 1259 and 669. A color bar on the right indicates counts from 0 to 1200.</p><table><thead><tr><th></th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1259</td><td>62</td></tr><tr><th>1</th><td>75</td><td>669</td></tr></tbody></table></div>		0	1	0	1259	62	1	75	669
	0	1	2	3																																
0	105	25	4	65																																
1	4	681	4	55																																
2	3	6	72	42																																
3	17	53	6	923																																
	0	1																																		
0	1259	62																																		
1	75	669																																		

As for models used to predict penalty, I similarly found that ensemble methods largely had superior performance. When predicting penalty type, Random Forest and Gradient Boost outperformed the other models. After hypertuning, Random Forest had an F1 Score (Weighted) of 0.923 and Gradient Boost a score of 0.906. Assessment of the Random Forest Model on the testing set resulted in a performance of 0.911 with acceptable performance on all classes (as visible through the classification report and confusion matrix). When examining feature importance, the model appears to place greater emphasis

on the officer's career length as well as the location and time of the incident. Media Coverage was not in the top 20 most important features.

Similar trends were observed in the model used to predict Penalty Count, with the Random Forest Model performing the best, with a F1 Score (weighted) of 0.96 when evaluated on the testing set. Examination of the confusion matrix suggests better performance at predicting single penalties but acceptable performance on the rest. Like the Penalty type model, the officer tenure and time/location of the incident appear to play a greater role. Media coverage appeared to have minimal role in number of penalties. I report the confusion matrix and performance metrics for the two penalty models below:

Table 5. NYPD Penalty Models Performance

Model	Type	Count																																																																																
F1 Score (Weighted)	0.911	0.967																																																																																
Confusion Matrix	<div><p>Confusion Matrix</p><table><tr><th>True \ Predicted</th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th></tr><tr><th>0</th><td>102</td><td>0</td><td>8</td><td>2</td><td>0</td><td>0</td><td>0</td></tr><tr><th>1</th><td>1</td><td>27</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><th>2</th><td>3</td><td>0</td><td>126</td><td>3</td><td>1</td><td>0</td><td>1</td></tr><tr><th>3</th><td>3</td><td>0</td><td>8</td><td>73</td><td>1</td><td>0</td><td>1</td></tr><tr><th>4</th><td>0</td><td>0</td><td>3</td><td>0</td><td>71</td><td>0</td><td>0</td></tr><tr><th>5</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>12</td><td>0</td></tr><tr><th>6</th><td>7</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>52</td></tr></table></div>	True \ Predicted	0	1	2	3	4	5	6	0	102	0	8	2	0	0	0	1	1	27	1	0	0	0	0	2	3	0	126	3	1	0	1	3	3	0	8	73	1	0	1	4	0	0	3	0	71	0	0	5	0	0	0	0	0	12	0	6	7	0	1	0	1	0	52	<div><p>Confusion Matrix</p><table><tr><th>True \ Predicted</th><th>0</th><th>1</th><th>2</th></tr><tr><th>0</th><td>61</td><td>13</td><td>0</td></tr><tr><th>1</th><td>0</td><td>419</td><td>0</td></tr><tr><th>2</th><td>0</td><td>3</td><td>12</td></tr></table></div>	True \ Predicted	0	1	2	0	61	13	0	1	0	419	0	2	0	3	12
True \ Predicted	0	1	2	3	4	5	6																																																																											
0	102	0	8	2	0	0	0																																																																											
1	1	27	1	0	0	0	0																																																																											
2	3	0	126	3	1	0	1																																																																											
3	3	0	8	73	1	0	1																																																																											
4	0	0	3	0	71	0	0																																																																											
5	0	0	0	0	0	12	0																																																																											
6	7	0	1	0	1	0	52																																																																											
True \ Predicted	0	1	2																																																																															
0	61	13	0																																																																															
1	0	419	0																																																																															
2	0	3	12																																																																															

Limitations and Challenges

There are some limitations given preprocessing choices and model selection. The biggest limitation is the imbalance in the dataset. While I have attempted to address it using SMOTE, using SMOTE can increase the risk of overfitting. Simultaneously, given that SMOTE is generating synthetic new data points based on existing data, it might also potentially increase bias. In the dataset, the target variable 'CCRB Complaint Disposition' had a relatively even distribution of samples for 'Substantiated' and 'Unsubstantiated' classes but lower numbers for 'Exonerated' and 'Unfounded.' Based on confusion matrix results, it appears that SMOTE was not able to fully address the imbalance. Models were generally better at predicting the majority classes and performed poorly on minority classes. Dichotomizing the variable improved performance though that model appears more prone to False Negatives, something not ideal given the context of the model.

As for NYPD Officer Penalty, SMOTE appears to have been more successful. Model performance was acceptable for all penalty types, though I do observe better performing at predicting single penalties in the count model. However, the preprocessing approach utilized has its limitations. In this dataset, officers may receive multiple penalties or singular penalties. When considering this issue, I initially built models to predict the likelihood of each penalty independently. However, this approach doesn't take into account other penalties that are also applied and assumes that each penalty is independent to the others. As such, that approach would have resulted in models that don't account for co-occurring penalties.

To avoid this issue, I instead fitted two models: one that predicted the number of penalties and another that predicted penalty type. For the latter, I recoded multiple penalties to the most severe penalty. I effectively thus developed a model that predicted the most severe penalty type that would be given, while also developing another model to assess whether it would be the only penalty given. This approach allowed me to avoid assuming penalty independence while still allowing me to tackle the question at hand. However, ideally, a multilabel classification approach should be utilized.

Other challenges include the computational cost. In this case, ensemble method models had the highest performance, which wasn't unexpected. Hyperparameter tuning was too costly to perform via grid search so I had to opt for Random Search. Other limitations include the issue of (1) temporality and (2) complaint clustering. Currently, these models include variables that account for past cases on an individual officer-level. However, this assumes that community-wide events (i.e. a large anti-cop movement or protest) or peer-level events (i.e. a colleague's misconduct case) has no effect. While we can account for the latter to some degree by including a variable for precinct, the temporality of such events and other spillover effects are not account for. Simultaneously, it is unclear how the clustering of multiple allegations under singular complaints is affecting results. For example, does one officer's allegation under a share complaint affect another officer? Similarly, do co-occurring complaints from various officers affect the other if they occur around the same time period?

Reflections

Using the data, I was able to construct three high-performing models that can effectively predict case outcomes and penalties allotted. While feature importance evaluation suggested that media coverage does not play a significant role, it also pointed out several factors that do. These understandings can help inform policies and highlight potential biases that occur in misconduct cases. Similarly, it can help inform legal action by allowing complaints to understand potential likelihoods as well as evidence to leverage in cases.

While this project can be refined (i.e. multilabel classification, expanded dataset, better collection of media coverage data), it ultimately completed its goals to a satisfactory degree. Future work should focus on refining the models to better capture multiple penalties as well as increase focus on the issue of temporality.

References

- About NYPD*. (n.d.). Retrieved January 30, 2025, from <https://www.nyc.gov/site/nypd/about/about-nypd/about-nypd-landing.page>
- Chermak, S., McGarrell, E., & Gruenewald, J. (2006). Media coverage of police misconduct and attitudes toward police. *Policing: An International Journal of Police Strategies & Management*, 29(2), 261–281. <https://doi.org/10.1108/13639510610667664>
- Civilian Complaint Review Board: Allegations Against Police Officers | NYC Open Data*. (n.d.). Retrieved January 31, 2025, from https://data.cityofnewyork.us/Public-Safety/Civilian-Complaint-Review-Board-Allegations-Against/6xgr-kwjq/about_data
- Civilian Complaint Review Board (CCRB) Database*. (n.d.). Retrieved January 31, 2025, from https://data.cityofnewyork.us/browse?Data-Collection_Data-Collection=CCRB%20Complaints%20Database
- Civilian Complaint Review Board: Complaints Against Police Officers | NYC Open Data*. (n.d.). Retrieved January 31, 2025, from https://data.cityofnewyork.us/Public-Safety/Civilian-Complaint-Review-Board-Complaints-Against/2mby-ccnw/about_data
- Civilian Complaint Review Board: Penalties | NYC Open Data*. (n.d.). Retrieved January 31, 2025, from https://data.cityofnewyork.us/Public-Safety/Civilian-Complaint-Review-Board-Penalties/keep-pkmh/about_data
- Crime Stats—Historical—NYPD*. (n.d.). Retrieved March 2, 2025, from <https://www.nyc.gov/site/nypd/stats/crime-statistics/historical.page>
- Cubitt, T. I. C., Gaub, J. E., & Holtfreter, K. (2022). Gender differences in serious police misconduct: A machine-learning analysis of the New York Police Department (NYPD). *Journal of Criminal Justice*, 82, 101976. <https://doi.org/10.1016/j.jcrimjus.2022.101976>
- Dowler, K., & Zawilski, V. (2007). Public perceptions of police misconduct and discrimination: Examining the impact of media consumption. *Journal of Criminal Justice*, 35(2), 193–203. <https://doi.org/10.1016/j.jcrimjus.2007.01.006>
- Harris, C. J., & Worden, R. E. (2014). The Effect of Sanctions on Police Misconduct. *Crime & Delinquency*, 60(8), 1258–1288. <https://doi.org/10.1177/0011128712466933>
- Headley, A. M., D'Alessio, S. J., & Stolzenberg, L. (2020). The Effect of a Complainant's Race and Ethnicity on Dispositional Outcome in Police Misconduct Cases in Chicago. *Race and Justice*, 10(1), 43–61. <https://doi.org/10.1177/2153368717726829>
- Mapping Police Violence*. (n.d.-a). Mapping Police Violence. Retrieved January 31, 2025, from <https://mappingpoliceviolence.org/methodology>
- Mapping Police Violence*. (n.d.-b). Airtable. Retrieved January 31, 2025, from <https://airtable.com/appzVzSeINK1S3EVR/shroOenW19l1m3w0H/tblxearKzw8W7ViN8>
- Rozema, K., & Schanzenbach, M. (2019). Good Cop, Bad Cop: Using Civilian Allegations to Predict Police Misconduct. *American Economic Journal: Economic Policy*, 11(2), 225–268.
- Tapp, S. N., & Davis, E. J. (2020). *Contacts Between Police and the Public, 2020*. US Department of Justice Office of Justice Programs Bureau of Justice Statistics.

Wright II, J. E. (2020). Will They Even Hear Me? How Race Influences Citizen Complaint Outcomes. *Public Performance & Management Review*, 43(2), 257–277.
<https://doi.org/10.1080/15309576.2019.1660188>