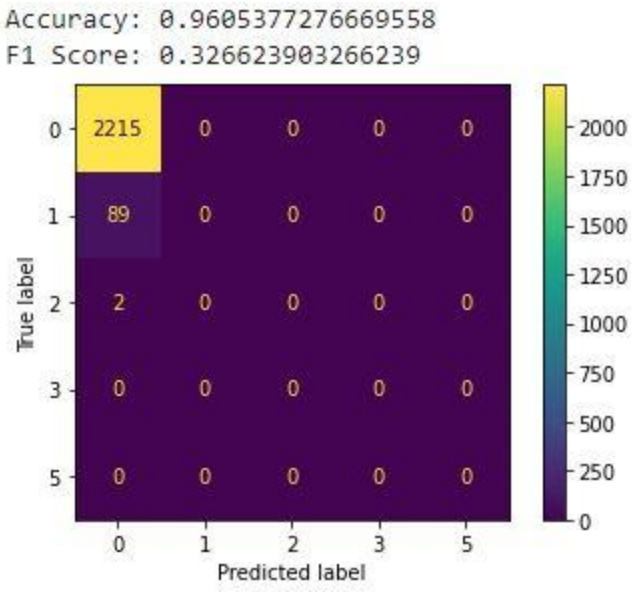
CP4 Findings

1) We are interested in forecasting future complaints for individual officers with a general forecasting model. We predict some of the features we will use to be the ones previously highlighted above, such as area of patrol, prior number of complaints, officer type, among others. This would be done with a forecasting model where the output is some likelihood of an officer having a future complaint filed against them within the time step we select.

We want to build upon our first visualization in CP3 and try to predict the likelihood of an officer being involved in an allegation in the future year. At first, we wanted to do a time series prediction based solely on allegations over the years, however, we decided against it because we really want to utilize our previous findings and to dig deeper on our established intuitions. We start with our queried officer data, including gender, race, age, years of service, allegation count, salary, and more, and we add the number of allegations each active officer received on a per year basis, starting from 1989. Data from before 1989 are excluded due to the extremely low reports and the lack of continuity between the years because of low sample size. We want to train our model by predicting the latest year in the data, 2018, using all other features described. We used a Support Vector Classifier as our learning model which uses a support vector machine to supervise learning.



Our model achieved an accuracy of 96 percent, but this is largely meaningless, due to the lack of consistency and variety in the data. Because most allegations in the 2018 column are 0, as is the case with any other years due to the allegations being generally spread out with little to no patterns, the model outputted mostly zeros. However, this really does not suggest that model has learned anything, as the F1 score has suggested. A F1 score of 0.31 is quite low and suggests that our model is inadequate at accurately forecasting future allegations. We consider this experiment to be quite lacking, and could certainly use more improvement. For one, we could try to increase the time steps we take between data entries, that way we might be able to get more generalized data patterns of allegations over the years, which could reveal more than our current model. We could also be more careful during feature selections, or perhaps go ahead with a time-series prediction to see if it would yield more interesting results.

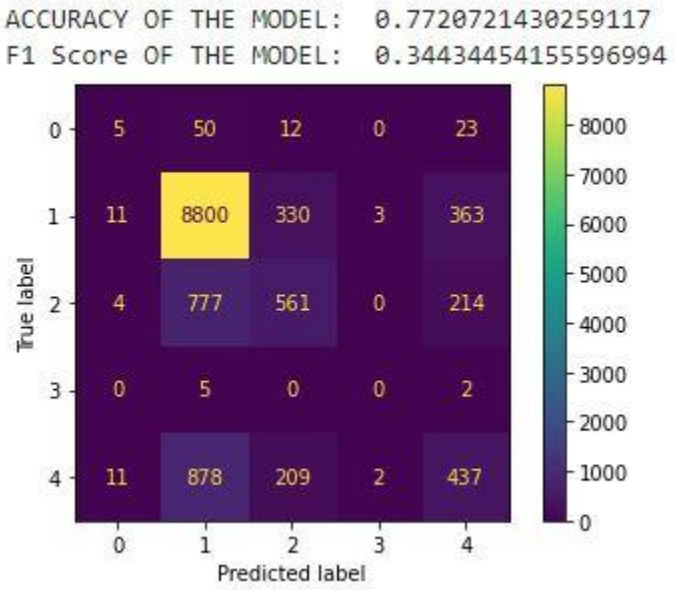
As for qualitative discoveries, we were unable to find many meaningful discoveries because our class-imbalance was so unclear. However, looking at the

confusion matrix, there around 5% of officers who had an allegation made against them that the model could not predict (not that it predicted many in the first place) — this could imply that there are only bad apples in the police force and that there is an unpredictability or culture of leniency to how and why officers have allegations made against them.

2) Since a large portion of our data discovery and visualizations also have to do with patrol demographics of the officers we are making predictions for, we are interested in modeling the disparity of severity and number of complaints made in predominantly Black and Brown neighborhoods compared to predominantly white ones. More specifically, this would look something similar to the following construction: if a new complaint X is made against officer Y, what is the likelihood that this complaint X is found in demographics A or B, where A is majority POC and B is not. We can do this with a simple Bayes model, where given features of complaint X and officer Y (things such as race of officer, number of priors, gender of complaintaint, formalized as [a,b,...z]), E[race=A]|[a,b,...y,z].

We were interested in being able to predict the race of a victim when a complaint is made against a specific officer. As we have explored a lot of data surrounding patrol demographics through the police beats profiles and average civilian allegations in each beat in our CP3, we can further expand upon our discovery by setting up a prediction model to test the racial biases within the Chicago police force. Using our queried data on victims of allegations and officers involved in the complaints, we are able to match

up and merge the associated victims and officers involved in an allegation using allegation IDs. We ended up with 63484 different matching victim/officer allegation profiles. Once we have done our data pre-processing, we can start on training a prediction model to predict the race of the alleged victims using sk-learn. We decided on using Random Forest Classifier as our learning model for classification, along with a 20/80 test-train split.



Using a basic model without a lot of hyperparameter tuning and optimization, our model showed a high accuracy with a low macro f1-score. We tested with absolute predictions though our problem is multi-class, and our model is inherently multi-class, the random forest classifier will take the label with the highest probability, and this makes testing significantly more clear (though, in the spirit of the original question, our model can also output the probabilities of each class, as shown in the last cell of the

notebook). As explained previously, a low macro f1 score weighs the recall and precision among all classes the same, thus a low macro f1 score with high accuracy implies that the majority class is predicted with good precision and recall, but the minority classes performed much more poorly. As seen above in the confusion matrix, the classes 1, 2, and 4 (being race= Black, Hispanic, White) are predicted with decent accuracy (all well above chance), but predictions for race=Black for the victim heavily outperformed the other predictions. This is likely due to the class imbalance, as a vast majority of the victims in the allegations made are Black. Improvements can be made on the model, but severe class imbalances are still a difficult to tackle problem with unclear techniques and approaches.

Our feature importances result is even more interesting; the beat id data turns out to be the most important feature in the prediction with a score of 0.31, way higher than any other features in our training model. This is very consistent with our findings from CP1, where we discovered that certain beats and locations with an extremely high POC population are clear outliers when it comes to the number of allegations with shockingly high numbers. Other features that are relevant but less important include honorable mentions percentiles and civilian allegations percentiles, these make sense as they are relative percentages thus offering a more holistic evaluations of the officers. Number of officers present/accused also seems relevant, most likely due to the fact that more officers involved correlates with a potentially more violent encounter. What was surprising to us was that the race and gender of the officers involved show little to no correlation or importance to our prediction. This indicates that career identity as an

officer is more influential than racial or gender identity when it comes to police force discrimination.

So what does this tell us in a qualitative sense? Looking at the feature importance can explain some of it, but these results imply that for a given officer, they likely will carry a racial bias in incidents where an allegation is made against them. Remember, these results are per officer; our test data only includes real officers and their real data. Secondly, the demographic policed definitely matters – perhaps officers have a higher likelihood for an abuse of power (thus translating to an allegation being made against them) in POC demographics, as the feature importance of the beat implies.