Checkpoint4: Machine Learning

Team members:
Hong Hong (NetID: hha7337)
Hui Ye (NetID:hyr1096)

Introduction

Our group attempts to conduct research on the diversity of police officers and discuss the relationship between policing diversity and misconduct rate, including race and gender representativeness. In this checkpoint, our group proposes two machine learning questions to gain a deeper understanding of the relationship between the diversity of police officers and misconduct rate.

It is worth noting that we made several fine adjustments to our current questions compared to our previous proposal. The questions in the proposal are more like models or directions for us to explore, we detail the methods and change a little on the question to make our machine learning more deep into the theme we want to explore. Throughout the experiments, we used the colab platform as well as sklearn for model training and prediction.

For the level definition in this analysis, you can refer to the data definition in our checkpoint1. In checkpoint1, we defined both the level of misconduct and the level of diversity by calculation and data segmentation.

Question1: What is the relationship between misconduct rate level and the diversity of the serving police?
To explore this problem, we will train the logistic regression models using both binary and multivariate classifications. Throughout the problem, we use an incremental improvement approach to optimize the model and explore the relationship between inputs and outputs.

Question2: What is the risk level of a police officer's civilian allegation rate?
In this question, we divide the level of civilian allegation percentile of a police officer into three categories: 0, 1, 2 respectively representing low, medium and high risk. We use the personal information of a police officer and the demographic information of the district they serve. The goal of this question is to predict whether a police officer may have misconduct, so as to be alert to the occurrence of this behavior.

The two questions are more like a step-by-step process of exploration, with the first question focusing on the relationship between the individual diversities of police officers, such as race, gender, and their regional misconduct rates, while the second question integrates the association between the groups and regions in which police officers work and their risk level on misconducting.

Question1

External link the colab:
https://colab.research.google.com/drive/1Qw4np_9mji468ysDq7Q7phXMDDtL-L66?usp=sharing

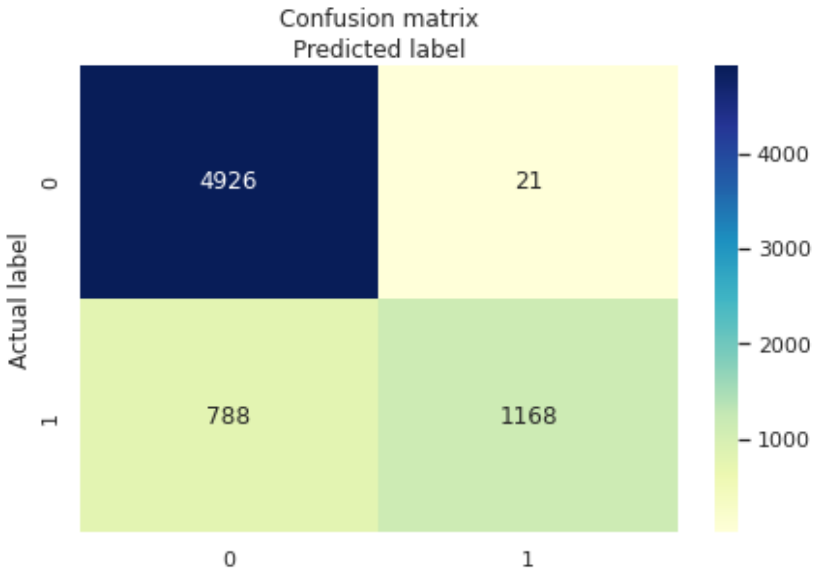
In this section, we first use sql for the extraction of the required data. Since the data used is not only the same 5 groups of districts defined in the previous checkpoint, we need to re-query the data. So we decided to correlate the data of 25 districts with the police officers who patrol and are on duty in them and select the column vector features about diversity in them.

Since we tried at the beginning to use only 5 groups of similar district groups and explore only two feature vectors, race, and gender, we observed that the data prediction was overfitted. After we adjusted the model to the parameters, we found that the results were not better, so we decided to select meaningful pairs of column vectors from the original database and input them into the model as feature vectors as well.

In the original scheme, we transformed the categorical data from the original data into binary categorical data based on the race, gender, and main population of the patrolled district of each police office, and used this as the input data, i.e., x. At the same time, we divided the misconduct rate of the corresponding district by medium if the misconduct rate is higher than 11%, we assume that the model considers the misconduct rate in this district to be relatively serious. According to this assumption, we also did the corresponding 0 1 binary treatment for the output y variable. However, the training results are not as good as expected.

We believe that this may be caused by various reasons, firstly the amount of data is not enough. Also the column feature vector is not enough, although there may be many columns in our input data, but in fact these columns are not mutually exclusive but interrelated, which means our input feature vector should be further optimized as well.

So we decided to first change the amount of input data and the column feature vectors. In the improved scheme, we added the police officer's major_award_count, trr_count, and unsustained_count as feature vectors, because we can reasonably speculate that these column feature vectors may also have some correlation with the misconduct rate. The amount of data is also increased to 10,353 records to better help the model training. Here we use a binary logistic regression model and divide the training set and test set in the ratio of 60% and 40%, and then make predictions to obtain the following results:

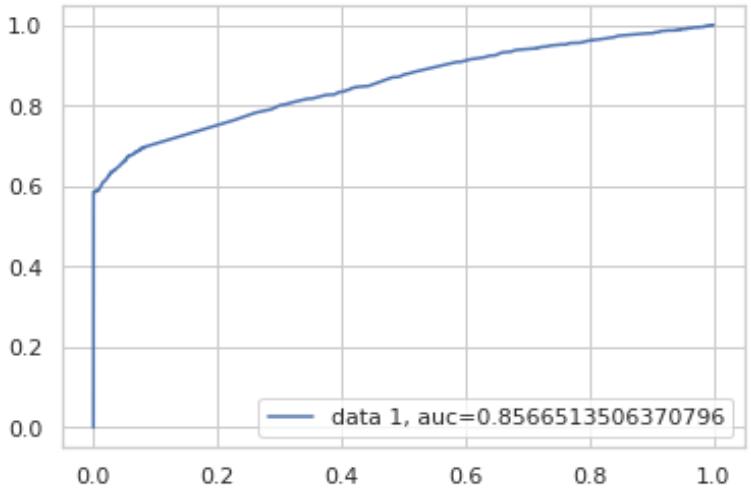



```
Accuracy: 0.882804577719832
Precision: 0.9823380992430614
Recall: 0.5971370143149284
```

Here, we use the confusion matrix to show the distribution of the predicted labels and the actual labels of our model. 0 represents the data with low misconduct rate level and 1 represents the data with medium and high misconduct rate level.

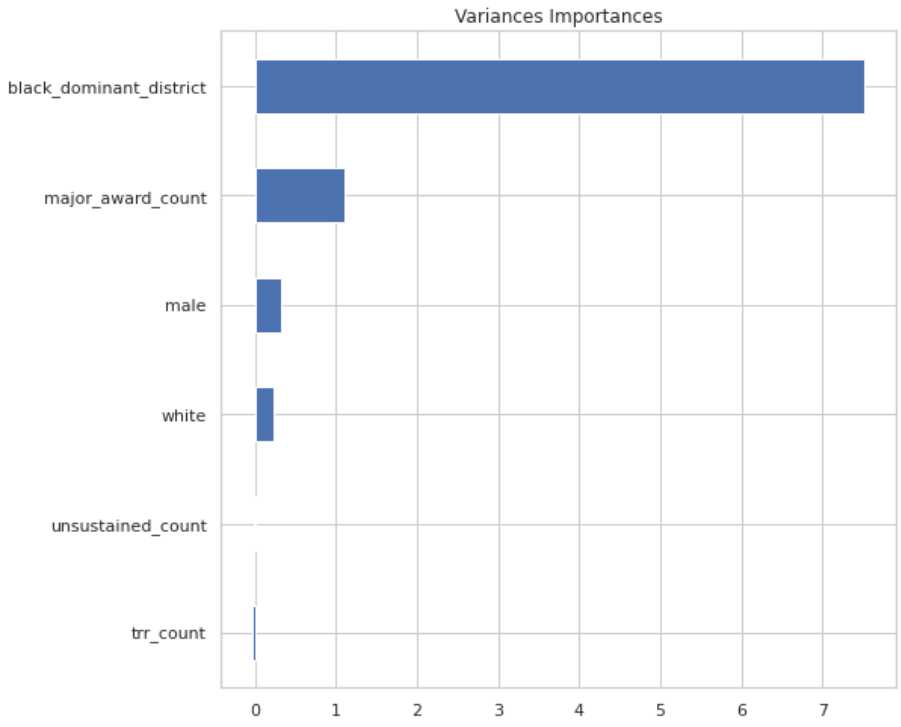
It can be seen that most of the labels are predicted to match the actual situation, and most of the predictions are wrong because the actual label is 1 and the prediction is 0. That is, the district's misconduct rate level is more serious, but in the prediction it is wrongly predicted to be low.

The prediction accuracy of the model is 88.3%, which can be called a relatively high model accuracy in machine learning, and this also represents that the binary logistic regression model we constructed has some effect and credibility.



Also, we evaluated the model using a ROC (Receiver Operation Characteristic) curve, roc curve is a performance measure for the classification problem under different threshold settings. ROC is the probability curve in the graph and AUC represents the degree or measure of separability. It tells how well the model is able to distinguish between different categories. The higher the AUC, the better the model is at predicting 0 to 0 and 1 to 1. By analogy, the higher the AUC, the higher the ability of the model to distinguish between misconduct rate as high or low. And our model has an AUC of 0.86, which indicates that our model has good prediction ability.

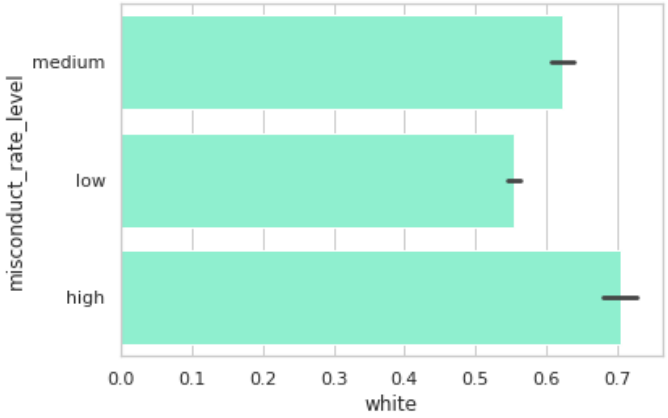
To further explore the effect of the column feature vectors of our model inputs on the output, we visualized the input variables as well as the output covariance. Rather surprisingly, whether or not a district is black-dominant district has a larger effect on the results for the misconduct rate for this district, which confirms what we have seen in class in articles related to this CPDB database. The graph below also shows that the race and gender of the police officer also have a significant effect on the misconduct rate of their district, with the degree of influence of gender being slightly larger than the degree of influence of race.



Then, to make the classification of the prediction results more detailed, we optimized our machine learning model so that its training and prediction can input and output multiple labels, i.e., a multiple logistic regression model. The input feature vector is treated the same as when using the binary model, while the output misconduct rate is labeled with a multivariate string, and we assign the label 'low' when the rate is less than 11%, and 'medium' when the rate is greater than or equal to 11% and less than 25%. when the rate is less than 11%, and 'medium' when the rate is greater than or equal to 11% and less than 25%. When the rate is greater than

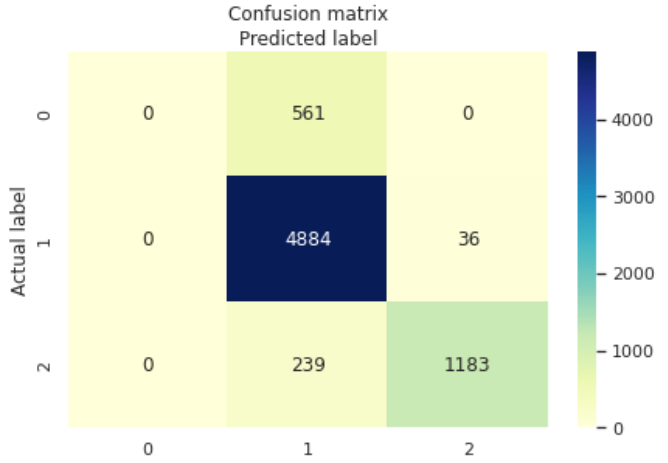
25%, we assign the label 'high' (refer to our checkpoint1 for the specific numerical classification criteria).

Before training the model, we did some overview of the data to have more insight into the data, from which we were able to find that the proportion of white police officers was the largest in areas with high misconduct rate, while the proportion of white police officers and the distribution of police officers of other ethnicities was nearly half in areas with low misconduct rate.



Similarly, we also output the confusion matrix and the model prediction accuracy to measure the validity of our model training：

In the confusion matrix, the labels 0, 1, and 2 represent the output y-labels 'low', 'medium', and 'high', respectively. From the figure below, we can observe that most of the predicted labels of the districts with the original labels of medium and high are consistent with the original ones, while the more anomalous case is that the predicted labels of the districts with the original labels of low are inaccurate and all of them are predicted to be medium labels. This may reflect that the input feature vectors of districts with higher misconduct rate have more obvious features that can be recognized by our model, while the input features of districts with low rate may be more similar to those of medium districts. In addition, the prediction accuracy of our model can still reach 87%, indicating that our multiple logistic regression model has some confidence in predicting districts with high misconduct rate.

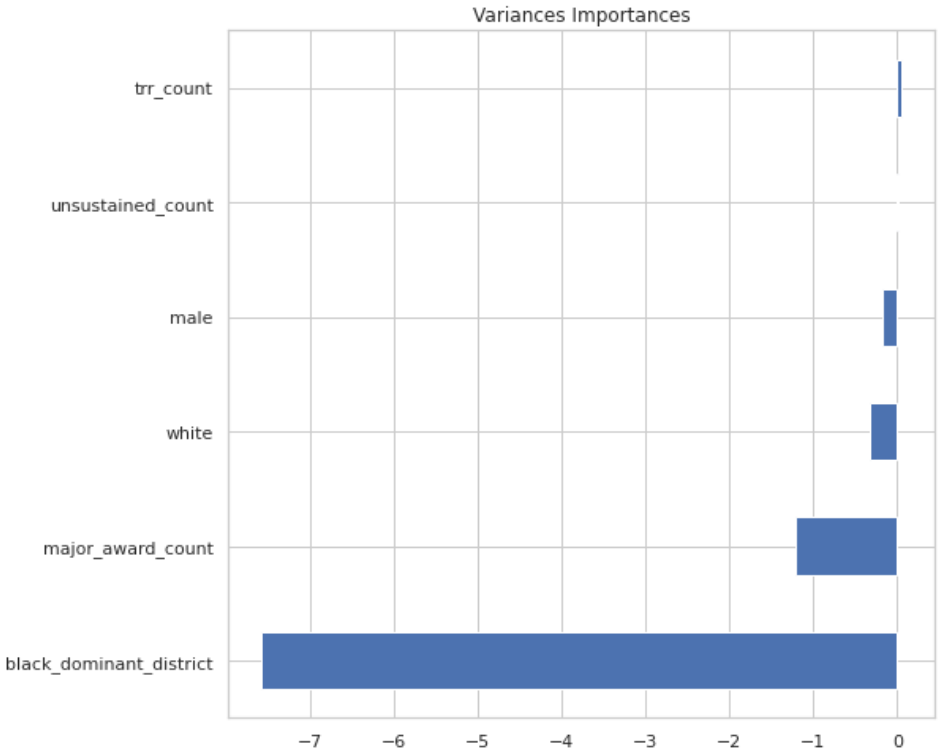


```
Accuracy: 0.8788932348254382
```

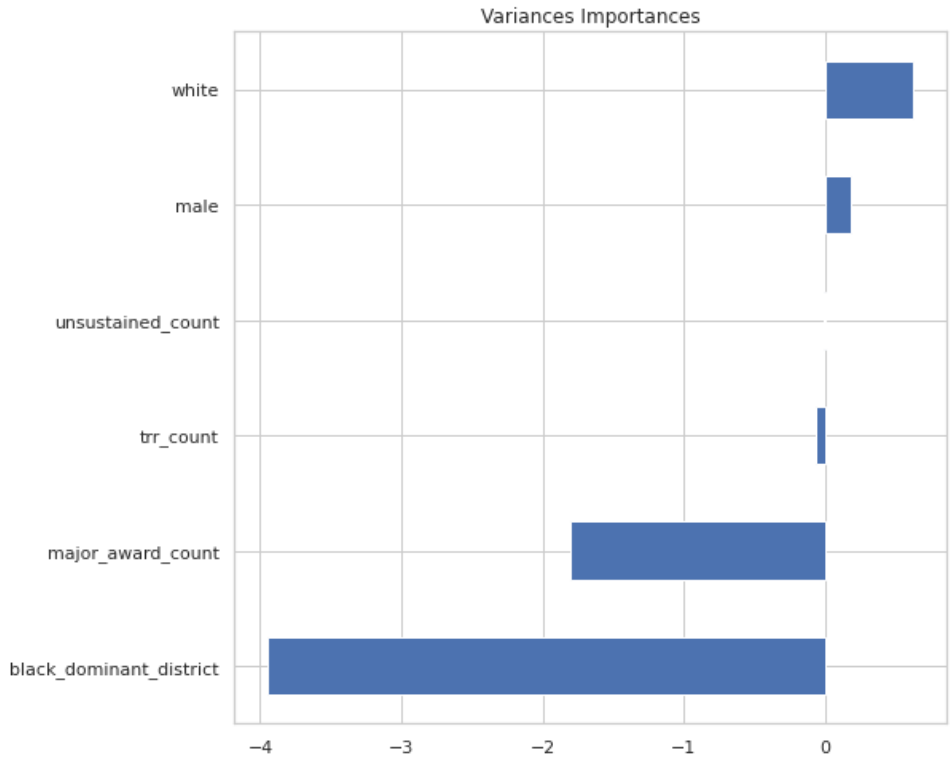
Since we used multivariate labels, I visualized the effect of the input feature vector on the different output labels by splitting the covariance.

```
print(logreg_high.coef_)
```

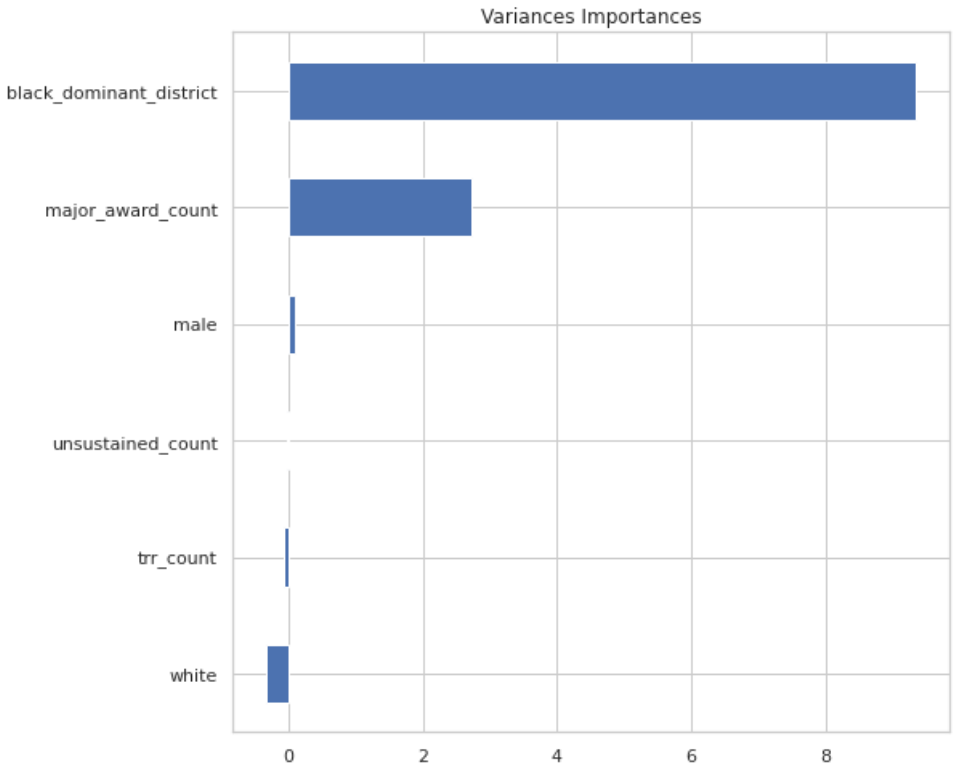
```
[[ 5.72071029e-01 -3.95765154e+00  1.96320879e-01 -1.85649645e+00
  -6.34984232e-02 -1.98429810e-03]
 [-2.54134894e-01 -7.47217424e+00 -2.04192644e-01 -1.05724230e+00
   4.62912708e-02  9.28920582e-03]
 [-4.30109835e-01  9.15856398e+00  1.75048611e-01  2.48403633e+00
  -4.18438736e-02 -1.65918700e-02]]
```



Misconduct_rate_level = ‘low’



Misconduct_rate = ‘medium’



Misconduct_rate = high

We can see that the misconduct rate of a district is related to the gender and race of police officers patrolling in the area, especially in the area of medium, the larger the percentage of male and white, the more likely the output is medium misconduct rate, while the low misconduct level area is just the opposite, if the area of If the police in the area of white accounted for less, or male police accounted for less, and if the area is not a predominantly black area, it is more likely to be a low misconduct level area. Conversely, observing high level areas, where police race and gender have an effect on their predictions, but where black-dominated neighborhoods have the most significant effect on predicted outcomes.

Question2

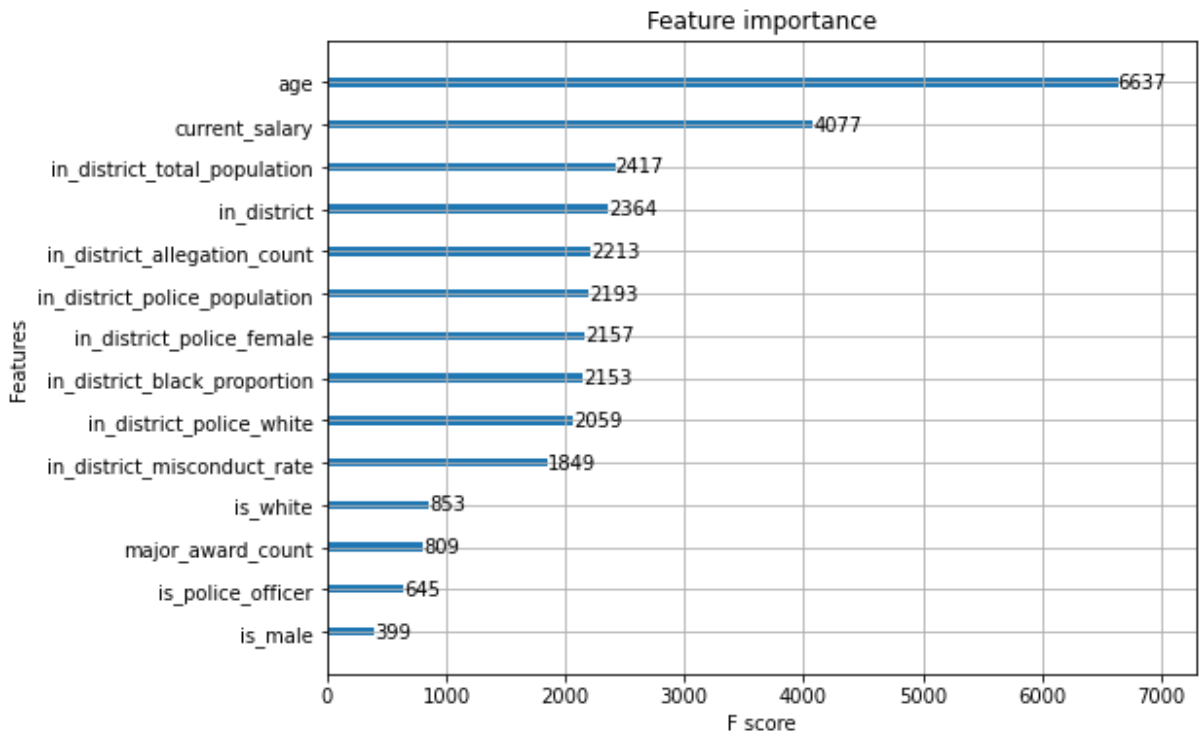
External link to colab:
https://colab.research.google.com/drive/19vbP4gJBUZSh7VpjllxTgygoXs-xEMRU?usp=sharing

In this section, we use OpenRefine to clean the data. The data includes 5,782 active police officers with comprehensive information about their age, race, gender, salary and award. The data also describes the demographic information of the district they serve, including total population, the black proportion of the district, and the number of allegations. At the same time, the information of the police unit serving this district is provided, such as the race and gender diversity of the police unit.

We use XGBoost to analyze predictions. The data is divided into training data and testing data. We use the training data (with multiple features) to predict the risk level of a police officer's civilian allegations, and use testing data to verify the accuracy of our model.

Features we are described as below:

- Is_male: if the officer is male.
- Is_white: if the officer is white.
- Is_police_officer: if officer has higher rank
- Age: the age of the officer.
- Civilian_allegation_risk
- Current_salary: the salary of the officer.
- Major_award_count: the number of awards that the officer won.
- In_district: which district the officer serves.
- In_district_total_population. the population of the serving district.
- In_district_black_proportion: the proportion of the black in the serving district
- In_district_allegation_count: the total number of allegations happening in the district.
- In_district_police_population: the number of police officers that served the district.
- In_district_misconduct_rate: the misconduct rate of the district.
- In_district_police_white: the proportion of white police in the district.
- In_district_police_female the proportion of female police in the district.



We use the plot_importance() method to depict a simple bar chart representing the importance of each feature in our dataset. Looking at the feature importances returned by XGBoost (as the above figure shown), we see that age dominates the other features, followed by current_salary. Then, the rest features in the top ten are all relevant to the district they serve. The population of the district is the third important feature in the bar chart, which means a police officer tends to have misconduct when serving a district with a large population. The allegation number of a district and the number of police units also show a relatively high score. Those features all show that the group that police officers work for has an impact on their behavior, including the district and police unit.

When we see the diversity feature, it's not hard to see that both the female proportion and white proportion of the police unit in a district have an effect on a police officer's misconduct. This is consistent with the findings we have in checkpoint3. We also find that whether a police officer is white or is male has less important than the race and gender proportion of the police unit they work for. It's surprising that the occurrence of misconduct behavior is more likely to be influenced by the group than the individual. The demographic of the work environment, the district and police unit they work for, tends to be a significant feature for the police misconduct.

	0	1	2	result	actual
0	0.053260	0.759383	0.187357	1.0	1
1	0.037466	0.634912	0.327622	1.0	0
2	0.041640	0.587718	0.370642	1.0	1
3	0.115372	0.603177	0.281451	1.0	0
4	0.197002	0.679482	0.123517	1.0	2
...	...	...	...	...	...
574	0.032669	0.560490	0.406841	1.0	2
575	0.046486	0.497159	0.456355	1.0	2
576	0.139092	0.742498	0.118410	1.0	0
577	0.025457	0.854432	0.120110	1.0	1
578	0.056410	0.737548	0.206042	1.0	1

```
Accuracy: 70.98%
```

We also train the XGboost model to predict the misconduct level of a police officer based on the feature described above. The model will produce the possibility of each level and we adopt the greatest possibility as the result of the predicted result. We test the model on 579 police data. The accuracy of the model is 70.89%.

For the future extension, we can involve more features in our model, such as the tenure, whether they have changed the police unit, the economic condition of the district they serve. In this checkpoint, we only study the area by district, we can segment Chicago in a smaller group, not only geographically but also on a society basis, which makes us involve more information about a police officer's work environment and improves the accuracy of the prediction model.

Conclusion

Overall, the percentage and metric is defined according to the data, the result may be a little less reliable, but we would say that the findings by machine learning somewhat support the question or give us interest to explore in this direction.

Specifically, from question 1, we can see from the results and analysis that our hypothesis is reasonable, that is, the race and gender of police officers have a certain influence on the misconduct rate of their duty areas, for example, the higher the proportion of male police officers, the higher the level of their misconduct rate is likely to be. The most significant finding is that the misconduct rate is higher in balck-dominant districts compared to other areas, which in turn is the most significant factor in the model prediction, which confirms what we have seen reported in class and is also consistent with the conclusions we have drawn in checkpoint 3.

From question 2, we were able to learn that police officers are often prone to misconduct in highly populated areas, while the police group or service area in which the officer works may also be a factor in their influence. We also found that whether a police officer is white or male is less important than the race and gender ratios of the police unit in which they work. Surprisingly, the occurrence of misconduct was more likely to be influenced by the group than by the individual.

In fact, there are many open questions to explore using machine learning. That is, what are some of the police groups that are prone to misconduct? Why are black-dominant districts prone to have high rates of misconduct? Is there a pattern in the temporal distribution of these misconducts? (e.g., is it mostly in the morning, afternoon, or evening?) I think these are some of the questions that are worth exploring.

For the future work, we could involve more features for model development. In this checkpoint, we only study the area by district, we can segment Chicago in a smaller group, not only geographically but also on a society basis.
We could also select some interesting data points to gain a deeper understanding, like there are several districts that have really high misconduct rates which are shown below as 41% and 83% respectively.

21	59458	6770	0.1139	274	0.4157
23	62781	13617	0.2169	259	0.8375