

Checkpoint 4: Machine Learning

Introduction & Data preparation

In this checkpoint, we will use the repeaters' data from 2014-2018. For each repeater in this time period, we keep the count of each officer's misconduct under each misconduct category. Therefore a sample data will look like this:

	total	Conduct Unbecoming (Off-Duty)	False Arrest	Illegal Search	Lockup Procedures	Operation/Personnel Violations	Use Of Force	Verbal Abuse
60	25.0	1.0	2.0	10.0	2.0	2.0	8.0	0.0
72	3.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
135	11.0	0.0	3.0	2.0	1.0	2.0	3.0	0.0
152	31.0	0.0	0.0	1.0	1.0	13.0	9.0	3.0
193	3.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0

The first column represents the officer ids, where subsequent columns give the total as well as subtotal of the 19 categories. We then convert all the subtotals to ratios between subtotal and total. The point is to make sure for each officer, now the subtotals will sum to 1 and we thus normalize the data as machine learning data input. The transformed data looks like this:

	total	Conduct Unbecoming (Off-Duty)	False Arrest	Illegal Search	Lockup Procedures	Operation/Personnel Violations	Use Of Force	Verbal Abuse
60	25.0	0.04	0.08	0.40	0.08	0.08	0.32	0.00
72	3.0	0.00	0.00	0.33	0.00	0.00	0.33	0.33
135	11.0	0.00	0.27	0.18	0.09	0.18	0.27	0.00
152	31.0	0.00	0.00	0.03	0.03	0.42	0.29	0.10
193	3.0	0.00	0.00	0.00	0.00	0.33	0.33	0.33

Methods

For this checkpoint, We use LinearRegression to predict the total and individual number of complaint reports. We recall that the four major misconduct categories are 'Use of Force', 'False Arrest', 'Illegal Search' and 'Operation/Personnel Violations'. We create a subset of the above data with these four categories as our 'X' data. We then use the total number as the 'Y' data. The goal is to, by training a LinearRegression model, calculate the coefficients and intercept so that the predicted value can be expressed as:

$$Y_{\text{hat}} = Y_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4. (1)$$

Results

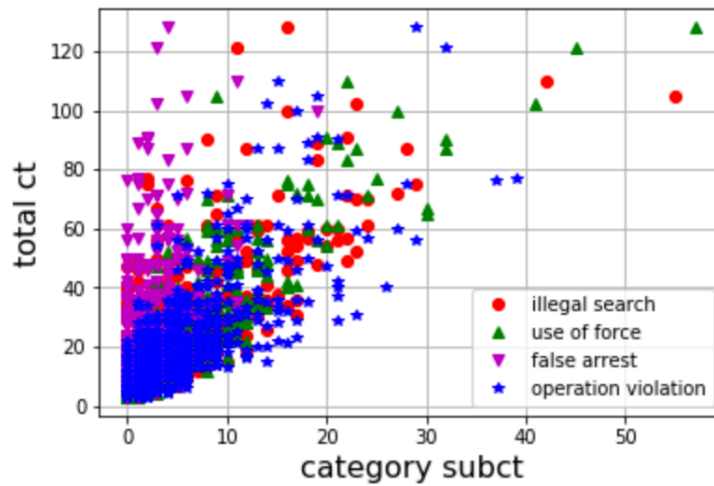


Figure 1: number of total reports v.s. number of report in each category

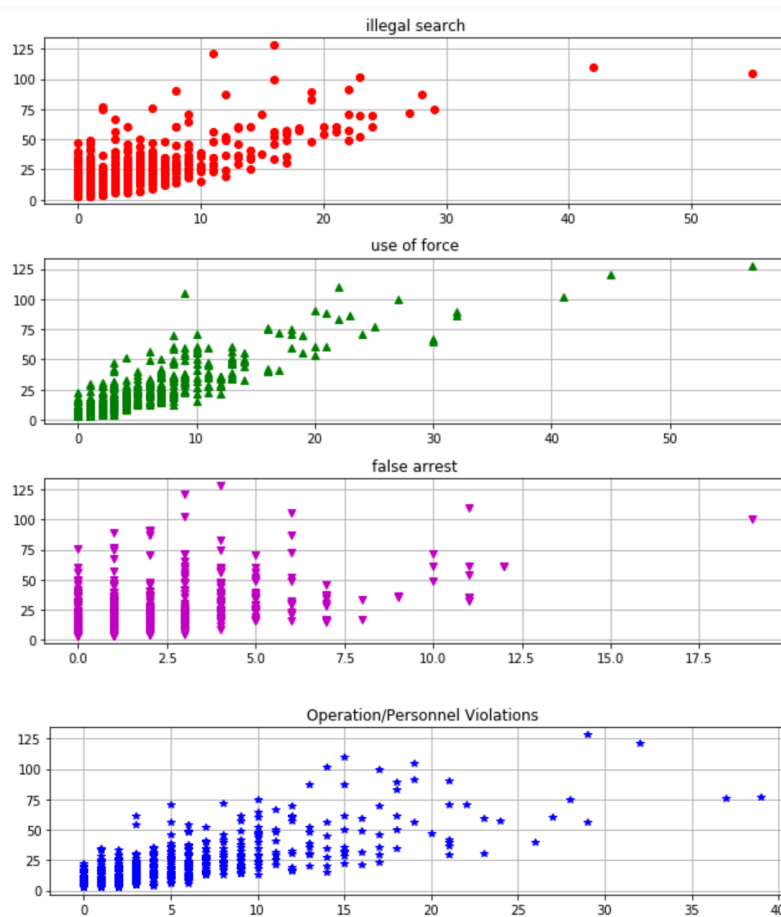


Figure 2: number of total reports v.s. number of report in each category(cont.)

From these two figures, we are confident that LinearRegression will give a reliable prediction on the total and individual number of complaints. When training the model, out of 557 repeaters we randomly select 80% of them as train set and use the rest as test set. The model is optimized by minimizing Mean Squared Error.

```

MSE = 9.016339529764586
RMSE = 3.0027220200618947
score for training set = 0.970712010615564
score for testing set = 0.9531618531757248

```

Figure 3: Evaluation of Model Performance

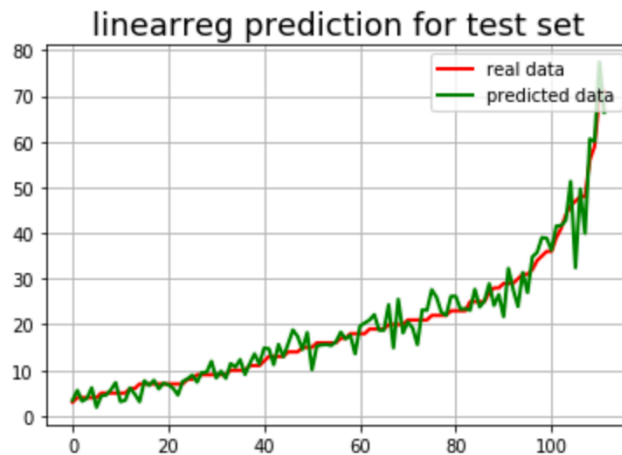


Figure 4: Evaluation of Model Performance (cont.)

Overall we can see that by minimizing the MSE, the model performs well in both training set and test set. We now can output the weights and intercept to produce \hat{Y} .

By applying Equation (1):

```

intercept: 0.7490838986652193
coef: [1.14035569 1.35575529 1.02469764 1.32925022]
actual total cases: 12707.0
predicted total cases: 13981.43614211605

```

Figure 5: Predicted total number of reports

	Illegal Search	Use Of Force	False Arrest	Operation/Personnel Violations
60	11.403557	10.846042	2.049395	2.658500
72	1.140356	1.355755	0.000000	0.000000
135	2.280711	4.067266	3.074093	2.658500
152	1.140356	12.201798	0.000000	17.280253
193	0.000000	1.355755	0.000000	1.329250
...
32400	2.280711	5.423021	5.123488	2.658500
32415	4.561423	1.355755	1.024698	1.329250
32423	5.701778	16.269063	1.024698	5.317001
32430	7.982490	4.067266	2.049395	2.658500
32435	7.982490	5.423021	2.049395	3.987751

Figure 6: Predicted individual number of reports

Now we can see that the predicted total number of cases will be 13981 and we can also print out the predicted individual number of cases.

Other Findings

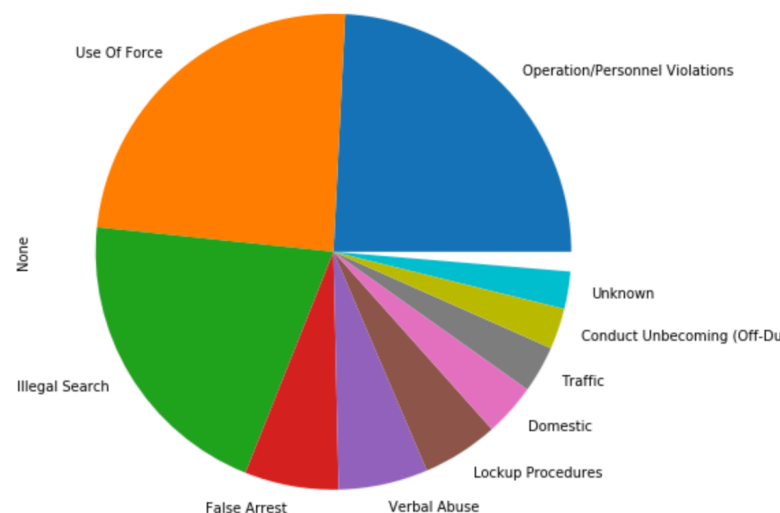
We conclude our findings with an in-depth look into the misconduct categories. We use K-Means Clustering as our machine learning model because we want to group repeaters of the same misconduct pattern into the same cluster.

We first use s-score to determine the optimal number of clusters and we get optimal K equals 2. Then we train the model and generate label (0 or 1) for each officer record and then split the dataframe according to the generated label.

Cluster 0 consists of 79 officers out of a total of 557 officers (14% of the population). The mean count of total allegations is 60.65 with a standard deviation of 20. Therefore the 95% confidence interval constructed is [56.235, 65.056]

A more detailed breakdown for the ratio of each category v.s all categories is as follows:

Conduct Unbecoming (Off-Duty)	0.027848
False Arrest	0.063797
Illegal Search	0.206076
Lockup Procedures	0.051646
Operation/Personnel Violations	0.242152
Use Of Force	0.241392
Verbal Abuse	0.060886
Criminal Misconduct	0.020380
Domestic	0.035823
Traffic	0.031772
Bribery / Official Corruption	0.002911
Supervisory Responsibilities	0.008481
Money / Property	0.004557
Medical	0.000253
Racial Profiling	0.000253
Excessive Force	0.000000
Drug / Alcohol Abuse	0.000759



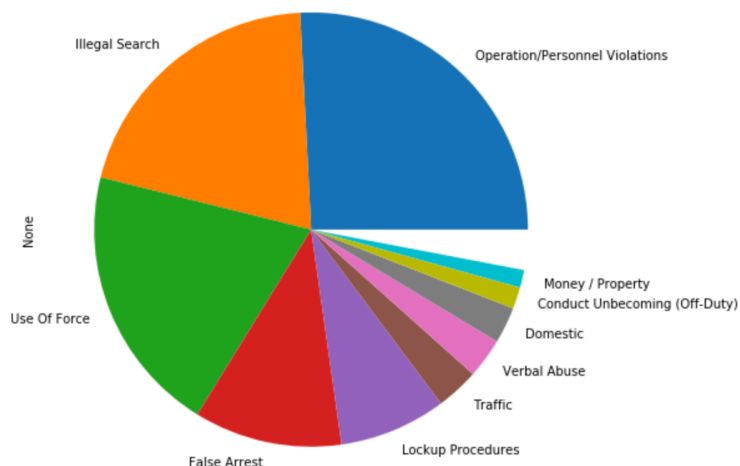
We can see that the major misconduct categories are:

- 1) Operations/Personnel Violations
- 2) Use of Force
- 3) Illegal Search
- 4) False Arrest
- 5) Verbal Abuse

Cluster 1 consists of 478 officers out of a total of 557 officers (86% of the population). The mean count of total allegations is 16.56 with a standard deviation of 9.40. Therefore the 95% confidence interval constructed is [15.718, 17.403]

A more detailed breakdown for the ratio of each category v.s all categories is as follows:

Conduct Unbecoming (Off-Duty)	0.016234
False Arrest	0.110439
Illegal Search	0.203138
Lockup Procedures	0.079728
Operation/Personnel Violations	0.258138
Use Of Force	0.200795
Verbal Abuse	0.029791
Criminal Misconduct	0.011130
Domestic	0.026987
Traffic	0.031841
Bribery / Official Corruption	0.005230
Supervisory Responsibilities	0.006820
Money / Property	0.013013
Medical	0.000377
Racial Profiling	0.001381
Excessive Force	0.001025
Drug / Alcohol Abuse	0.002762
Unknown	0.008368



We can see that the major misconduct categories are:

- 1) Operations/Personnel Violations
- 2) Illegal Search
- 3) Use of Force
- 4) False Arrest
- 5) Lockup Procedures

We can see that these two clusters, although differing in the average number of cases for repeater, have almost the same outlook for misconduct categories. This further corroborates our assumption that the repeaters have clear pattern on misconduct categories: 'Use of Force', 'False Arrest', 'Illegal Search' and 'Operation/Personnel Violations'. We need to more focus on these categories and this can be a potential research area for NLP by analyzing the details of complaint reports.