CPDP Report
By The Clever Eagles

Omkar Ranadive, Nikita Raut, Anupam Tripathi

Theme: The theme of our project was to understand the impact and analyze the trends after the CPDB went public. Our project compares and analyzes the trends before release of CPDB (before 2015) and compares it with the trends observed after release of CPDB (from 2015 onwards). We chose the year 2015 as the first subset of the dataset was released in 2015.

Index:

For each topic, we have only included our most significant findings in this report.

Checkpoint 1 - Relational Analytics

1. **Officers of which rank had the most complaints against them?**

In this question, we tried to investigate how do the number of complaints vary with the rank of an officer and how has this trend changed over the years. Here, we have taken the normalized count of complaints; that is, we divide the number of complaints for that rank by the total number of officers of that rank.

Output: (For the top 5 ranks with most complaints)

(Before 2015)

	normalized_count numeric	rank character varying (100)
1	94.00000000000000000000	Director Of Caps
2	22.00000000000000000000	Assistant Superintendent
3	10.94721407624633430472	Field Training Officer
4	10.34558823529411765368	Commander
5	10.16595135908440615340	Sergeant

(From 2015)

	normalized_count numeric	rank character varying (100)
1	0.46920821114369501440	Field Training Officer
2	0.20400572246065808014	Sergeant
3	0.19959940782025603065 2952	Police Officer
4	0.10312500000000000033	Lieutenant
5	0.08086077600260841240	Detective

Analysis:

We found out that higher ranked officers like Director of Caps and Assistant Superintendent had a high normalized count before 2015 but from 2015 onwards, they weren't in the list of top 5 officers.

One important observation to make is to look at the rank of Sergeant. The Sergeant rank has the second highest number of officer count (3495 officers) but still it appeared in the top 5 ranks in both before 2015 and from 2015 onwards. **So, studying the rank of sergeant could provide important insights. We explored this in checkpoint-4.**

2. **How many of the incidents have been carried out by off-duty cops and how has this trend changed after release of CPDB?**

We had observed that some of the most extreme incidents were carried out by off-duty officers. So, we wanted to observe how significant is this trend.

Output:

Before 2015: We found that before 2015 there were 1479 off-duty officers and 59066 on-duty officers. **Ratio of off-duty/on-duty = 0.025**

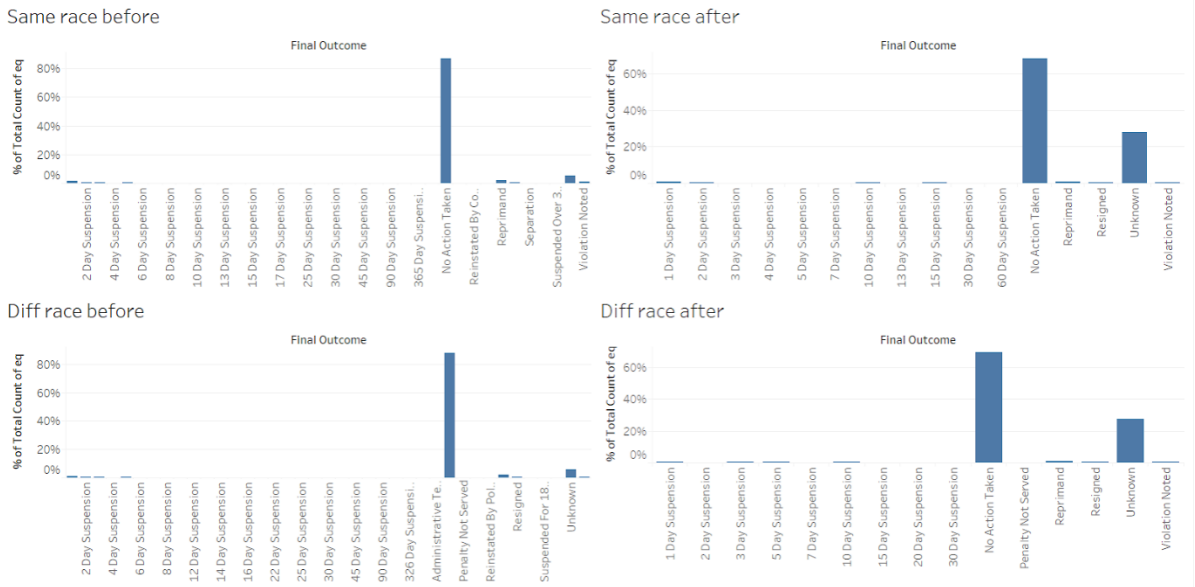
After 2015: After 2015 we found that there were 121 off-duty officers and 6353 on-duty officers. **Ratio = 0.019**

We can see that there is no significant difference between the ratios. So, we decided to go one step further and we segregated the off-duty officers by the beat-id.

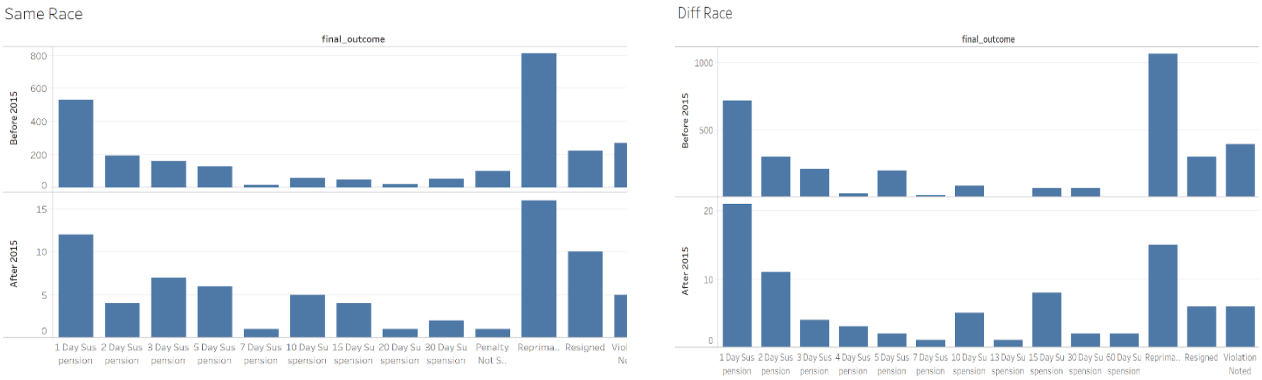
Output: Normalized count of off-duty officers by beat-id

As we can see, the relative distribution remains similar. But one interesting thing we observed was that there are certain beat ids where the count of off-duty cops is high in both the cases. So, it would be worthwhile to study these specific beat-ids.

3. Does the investigator's race affect the punishment given to the offending officer?



From the bar chart, we observed that the normalized outcome was similar regardless of whether the investigator's race was same or different to that of the officer. We went one step ahead and removed the "no action taken" and "unknown" columns from the bar chart.

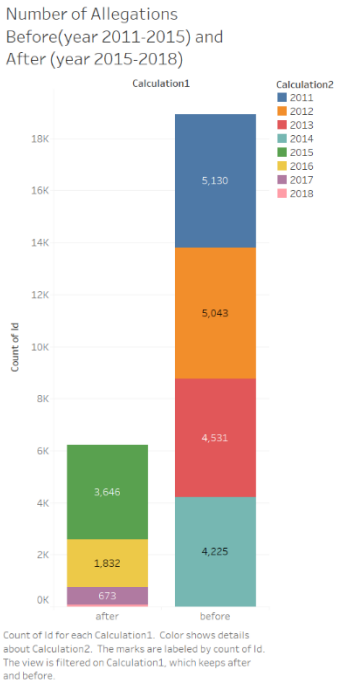


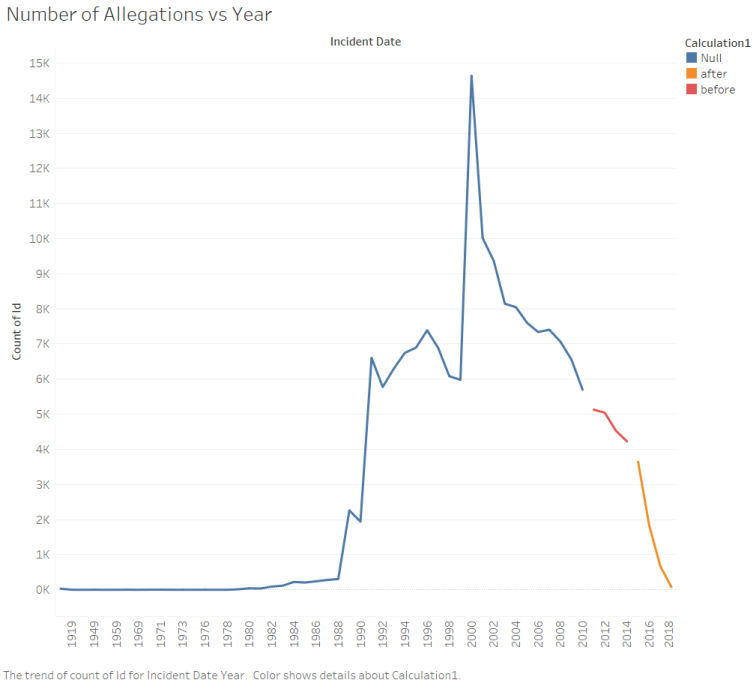
We found a positive trend after plotting these bar charts. After 2015, the amount of "severe" punishments has increased. We can see that punishments like "15, 30-day suspension" are higher after 2015.

Main takeaway: Based on our results of checkpoint 1, we have observed two important trends. One, a positive trend can clearly be observed after release of CPDB, second, there are certain factors (like off-duty cops racking up allegations in certain beat-ids) even after release of CPDB.

Checkpoint 2 - Visualizations

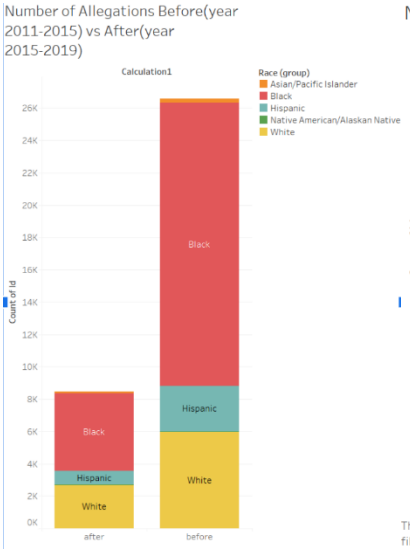
1. Visualizing whether there has been any decrease in allegations after the CPDP data has gone public with the help of line charts (Allegation vs Time) and histograms (Before vs After)

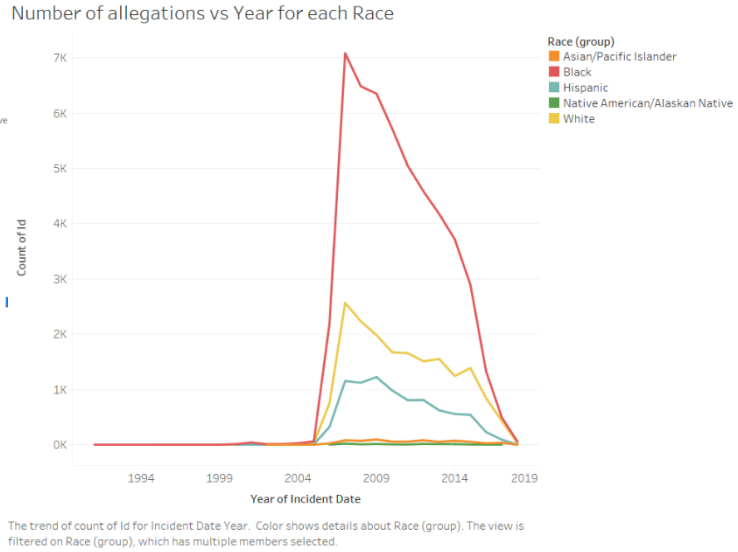




For the sake of fair comparison, we have compared the data from 4 years before 2015 and 4 years from 2015 onwards. We can clearly see a clean decrease after 2015.

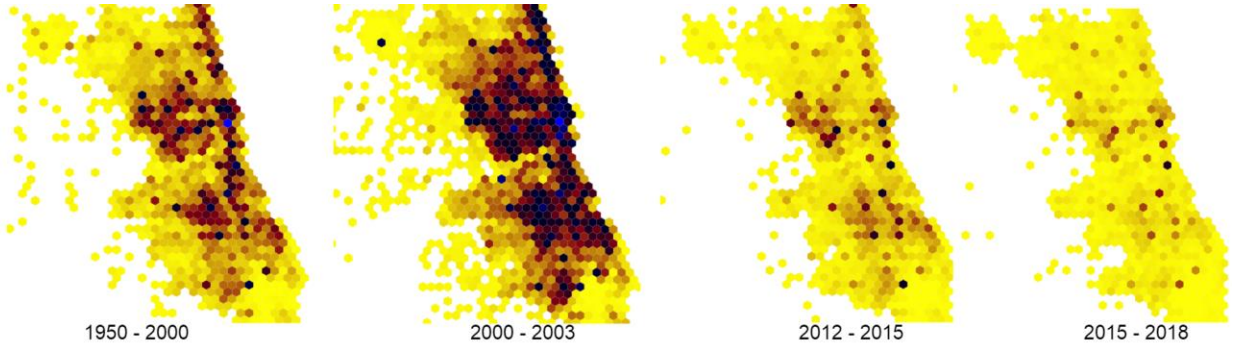
2. Visualizing trends in number of allegations segregated by the race of the victim and analyzing how these trends have changed over the years





We found a very interesting result. There was a significant drop in the number of allegations for all races in the year 2007. The second highest drop was in the year 2015. We posit that this is due to the following events: In 2007, IPRA got replaced by OPS. Moreover, the request for making officer data public on the Bond case was agreed in 2007. In 2015, the CPDB was released and also COPA replaced IPRA. From the graphs we can see that these events have had a significant positive impact.

3. Heat map of areas in which allegations have taken place over the years



There are some areas were the hexagons are very dark for most of the intervals of years signifying spots where these misconducts are very common. We can see that highly are targeted areas remain same even after CPDB went public. We also found some officers have been constant to some areas over the years eg: area with beat id 261 had officer Michael Clifton through the years 2006 - 2016, beat id 215 had officer Paul Major 2000-2013. **This also ties well with the result we found in Checkpoint-1 (beat-id is a significant factor for off-duty cop misconduct).**

Main takeaway: The results obtained in this checkpoint conformed with the results which we got in the previous one. We observed an overall positive trend but found that certain factors like beat-id are showing grim trends (darker areas) even after release of CPDB.

Checkpoint-3 Data Cleaning and Integration

1. What is the normalized settlement amount per officer rank and how has this trend changed after the release of CPDB?

Before 2015

1	Chief	10000.0000000000000000
2	First Deputy Superintendent	45000.000000000000
3	Sergeant	106285.183035714286
4	Deputy Chief	114700.000000000000
5	Police Officer	144259.818820224719
6	Commander	157500.000000000000
7	Field Training Officer	187967.106382978723
8	Lieutenant	314417.400000000000
9	Captain	403875.000000000000
10	Detective	465569.301075268817
11	Superintendent Of Police	1722778.333333333333

From 2015 onwards

1	Captain	10000.0000000000000000
2	Detective	24333.555555555556
3	Police Officer	30296.135922330097
4	Deputy Chief	45000.000000000000
5		45000.000000000000
6	Sergeant	53116.933333333333
7	Lieutenant	658437.500000000000
8	Field Training Officer	732522.750000000000

Before 2015, we observed that lower ranked officers have higher settlement average, which was contradictory to our expectations. After analyzing further, we found that cases against higher ranked officers were generally less severe which could explain the lower settlement amounts. Also, as seen above, Police Officers and Superintendent of Police are outliers. To analyze this further, we looked into the Superintendent of Police rank and found out that majority of the settlement amount was linked to Terry Hillard. **From 2015 onwards, we can see that the average settlement amount has gone down, which is a positive trend**.

2. What is the average duration of cases per officer rank and how has this trend changed after release of CPDB?

Before 2015

	rank character varying (100)	settlementavg double precision
1	First Deputy Superintendent	245
2	Field Training Officer	649.723404255319
3	Sergeant	698.609865470852
4	Lieutenant	749.4
5	Police Officer	826.495007132668
6	Deputy Chief	995.8
7	Chief	1011
8	Detective	1057.35359116022
9	Commander	1119.16666666667
10		1196.66666666667
11	Captain	1208.69230769231
12	Superintendent Of Police	1451

From 2015

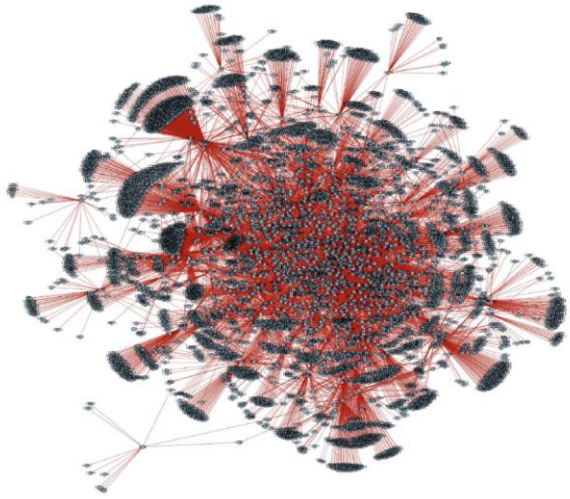
	rank character varying (100)	settlementavg double precision
1	Police Officer	221.089108910891
2	Field Training Officer	229.25
3		235
4	Captain	237
5	Lieutenant	308.25
6	Sergeant	350.7
7	Detective	352.444444444444
8	Deputy Chief	419

We again found a positive trend that from 2015, the cases seem to be settling faster.
Main takeaway: We found a positive trend in both the questions which again conformed with our previous observations.

Checkpoint 4 – Graph Analytics

1. Forming a co-accusal network where the officers are the nodes and they have an edge if they are co-accused in a case

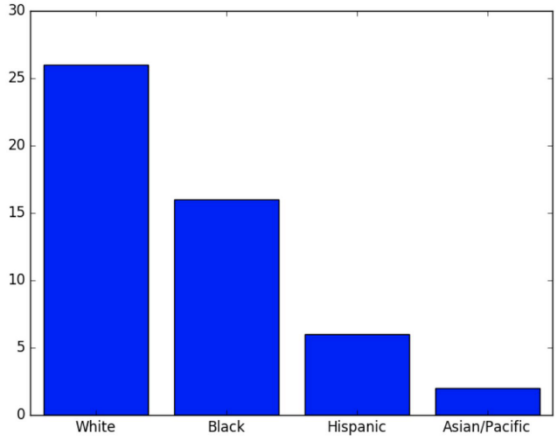
Before 2015

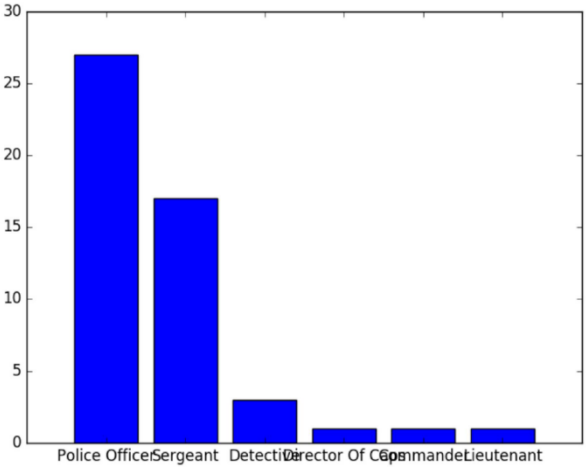


From 2015 onwards

To analyze the most important nodes, we used Page rank algorithm and found the top 50 most important officers.

Before 2015:



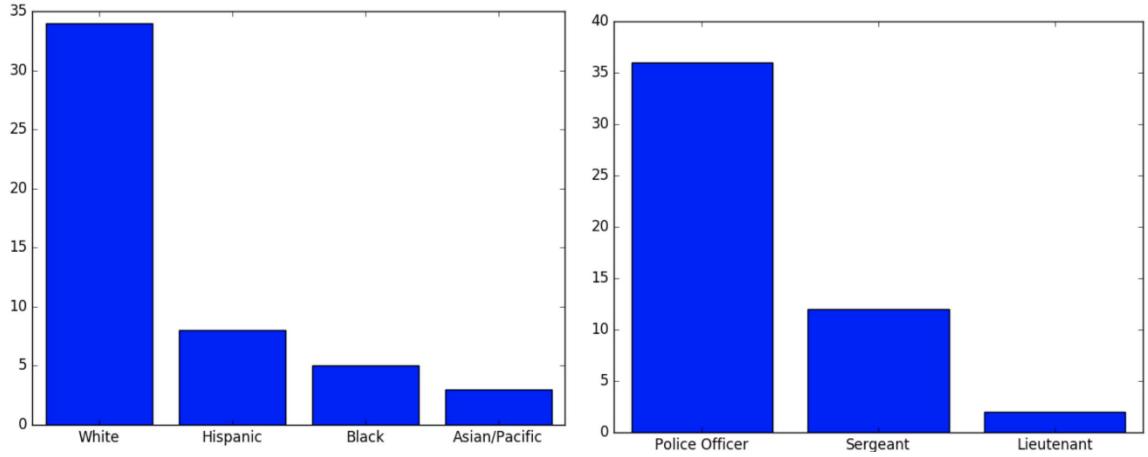


After applying page rank, we segregated the officers by race and rank. The following normalized counts were obtained: Asians = 0.67, Hispanic = 0.437, White = 0.553, Black = 0.441.

So, it was interesting to note that 2 out of the 3 officers were there in top 50 nodes and thus had the highest normalized count.

Similarly, the normalized results for ranks were as follows: Director of Caps = 1, Sergeant = 0.77, Police officer = 0.442, Detective = 0.375, Commander = 0.5

After 2015:



After applying page rank we found the following results for race: White = 0.72, Hispanic = 0.5, Black = 0.147, Asian = 1. **Yet again, we can see that the Asian officers are present in top 50 category.** For rank, we found the following results: Police Officer = 0.59, Sergeant = 0.54, Lieutenant = 0.6. **We can see that the rank of Sergeant has a fairly high normalized count in both cases. Also, in checkpoint-1 we found that the Sergeant rank had a high allegation rate. So we can say that the sergeant rank is an important one to study further.**

Main Takeaway: In this case, we found out that Asian officers and the rank of Sergeant were high up in normalized ranking, both before and after release of CPDB. This conforms with our second observation which we made in Checkpoint 1 – that there are some specific factors which have remained constant even after release of CPDB.

Checkpoint 5 – Machine Learning and Nature Language Processing

In this checkpoint, our focus was on analyzing how the distribution of data changed after release of CPDB.

1. In the first question, we answer the question about the data distribution by creating an ML model which flags officers as "bad cops" or "good cops". This is done by training the model on data before 2015 and then testing it on data from 2015 onwards.

Our hypothesize was that CPDB has changed the data distribution by introducing positive trends from 2015 onwards. Hence, the testing distribution should be different leading to low accuracy on the test set. We trained the dataset on two different classifiers: Here train accuracy is accuracy on data before 2015 and test accuracy is accuracy on data from 2015 onwards.

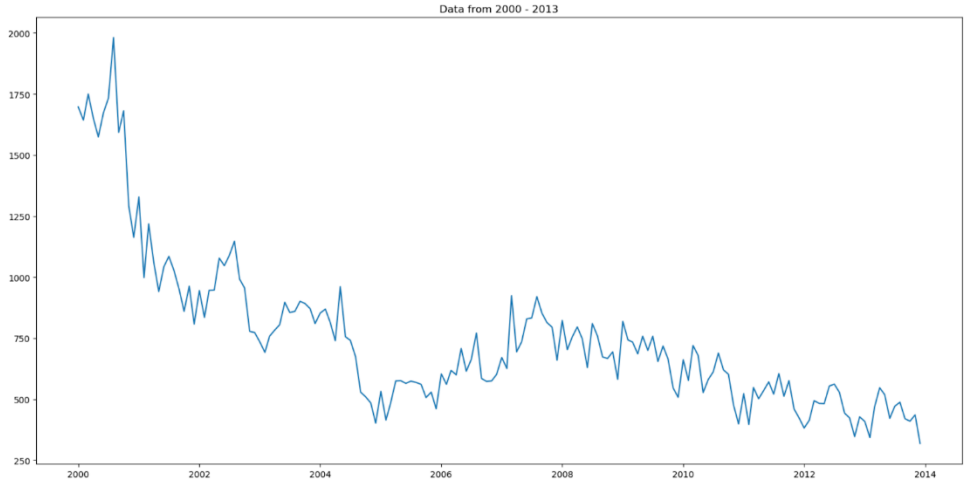
Results on Logistic Regression: Train accuracy: 0.6667, Test accuracy is: 0.1328

Results on SVM: Train accuracy: 0.6631, Test accuracy is: 0.1305

As we hypothesized, the test accuracy is significantly low as compared to the train accuracy. This is a strong indication that the test distribution is different from the train distribution and release of CPDB has had an impact on trends.

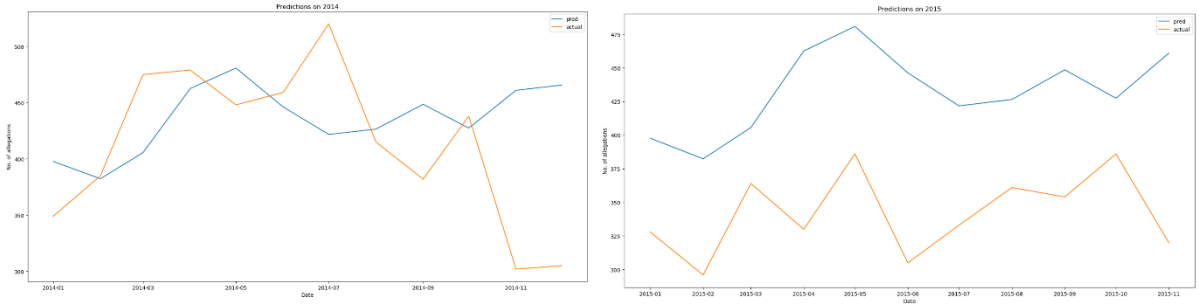
2. In the second question, we created an ML model which can predict the number of complaints in upcoming years. We trained the model on data before 2015 and then tested it on data from 2015 onwards.

Data distribution from 2000 to 2013:



Our model on data from 2014 and data from 2015:

Predictions on 2014 **Predictions on 2015**



We can clearly observe that our model was able to predict the data distribution from 2014 correctly but significantly deviated in predicting the distribution of 2015. This is a clear indication that from 2015 onwards, the trends have changed.

3. For this question, we group similar words in different categories from document text using the N-gram model and then try to analyze the differences in the groupings before 2015 and from 2015 onwards.

In this, we found out that due to the sparse nature of true labels in each category and the extremely noisy data (misspelled names, mislabeled data), the resulting groupings weren't great. Hence, it's difficult to perform analysis on the nature of distribution with the data available to us. Still, we found some interesting trends. Let's consider the category of nudity and penetration.

Before 2015: Words like 'clothes', 'pants', 'removed' were found. The word 'stomped' was strongly associated with the word 'removed' which kind of shows the general aggressive nature of the officers before 2015. There were also words like 'restaurant' and 'street' which gives us an idea of the kind of places were incidents of these category took place.

After 2015: Words like ‘parties’, ‘vehicles’ were found which shows the kind of places incidents of these categories took place.
Another interesting observation is that words like ‘drugs’, ‘cannabis’, ‘narcotics’ were present in groupings from both before 2015 and 2015 onwards showing that incidents of this category strongly overlap with incidents of drugs. Such trends can be studied for the different categories by observing the probability distributions given by our code.

4.Tagging documents by a “severity” measure
For this task, we hand engineered a list of words and assigned weightage to each word based on the average number of years of punishment associated with that crime. Fuzzy pattern matching was used to find matches. The most severe document we found had a score of 275. Then we divided the dataset into two parts – ‘Before 2015’ and ‘From 2015 onwards’. The average normalized severity measure before 2015 was 0.20713460389136065 and from 2015 onwards was 0.25810623556581985. Surprisingly, the average severity from 2015 onwards is higher than before 2015. Though we believe that the experiment should be rerun after cleaning the data better as at present it is very noisy.

Main takeaway: We found out that the data distribution significantly changed from 2015 onwards which is a clear indication of CPDB’s influence.

Conclusion

From all our checkpoints, we could make two main conclusions.

- There clearly has been a positive trend after the release of CPDB in 2015. This is backed up well with the help of the following observations – More sever punishments were given to officers from 2015 onwards (Checkpoint 1), there has been a decrease in the number of allegations and number of allegations by race (Checkpoint 2), Settlement Amount and average number of settlement days have gone down (Checkpoint 3) and most importantly the data distribution was found to be clearly different in Checkpoint 4. **All of these observations help us conclude that the release of CPDB has positively changed the trends**.
- There are certain factors which remain the same even after the release of CPDB and should be studied in more depth. This is backed up with the following observations – Off-duty cops had a high allegation count in specific beat-ids (Checkpoint 1), Certain areas remained dark (more incidents) over all the years (Checkpoint 2), Rank of Sergeant seems to be an important point in allegation graphs (Checkpoint 4), the few Asian officers are important nodes even after release of CPDB (Checkpoint 4).
  All of these observations help us conclude that certain factors have remained constant throughout and should be further scrutinized.