NY Daily Inmates in Custody

Springboard - Data Science - Capstone Two - Slidedeck

Explored and Written by Saleha Bakht

Overview

The dataset analyzed was provided by the City of New York on their Daily Inmates in Custody. While the City has not posed any questions with the release of its data, our exploration will look into exploring the patterns presented in the data to see if there are any unexpected findings. The goal is also to use machine learning processes to predict missing values and verify the findings in the dataframe.

Goals

1. Does race or gender have overly large influence on custody level?
2. Is there a pattern in infractions amongst certain groups of inmates?
3. Attempt a predictive model to see if the custody level of the inmate can be predicted.

Cleaning the Dataset

Cleaning the Dataset

- Replaced the original column names ‘BRADH’ and ‘SRG_FLG’ with ‘MENTAL_OBSERVATION’ and ‘GANG_AFFILIATION’ respectively.
- Dropped the columns ‘DISCHARGED_DT’ and ‘SEALED’.
- All rows that contained a single null value in the columns of ‘GENDER’, ‘RACE’, and ‘AGE’ were dropped.
- The rows that had null values in the ‘TOP_CHARGE’ column were imputed with ‘-1’.
- Broke up the time stamps in ‘ADMITTED_DT’ into columns for year, month, day, and hour.

Exploratory Data Analysis

Initial Findings

- There were 14.45 times as many male inmates as there were women.
- Low numbers of gang involvement.
- Proportionally higher numbers of women being under mental observation than men.
- More men had unrecorded values in their top charge but the proportion disparity is getting smaller and smaller by the year.
- In relevance to race, there was nothing noteworthy when comparing the genders in incarceration. Race had no bearing on the existence of infractions either. There are proportionally less of each race under mental observation except for the White and Islander populations. Those races have more people under mental observation than not.

Top Ten Most Common Charges for 18+ Y.O. Inmates

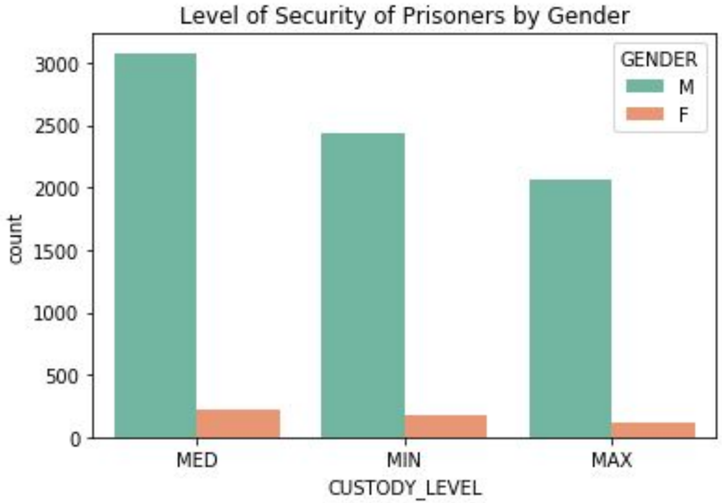
1. 125.25 is 'Murder in the second degree'.
2. 160.15 is 'Robbery in the first degree'.
3. 110-125.25, if treated as a range, is 'Attempt to commit a crime' to 'Murder in the first degree'. No singular offense was found for this charge.
4. 120.05 is 'Assault in the second degree'.
5. 265.03 is 'Criminal possession of a weapon in the second degree'.
6. 140.25 is 'Burglary in the second degree'.
7. 220.39 is 'Criminal sale of a controlled substance in the third degree'.
8. 220.16 is 'Criminal possession of a controlled substance in the third degree'.
9. 160.10 is 'Robbery in the second degree'.
10. 140.20 is 'Burglary in the third degree'.

Top Ten Most Common Charges for 16-17 Y.O. Inmates

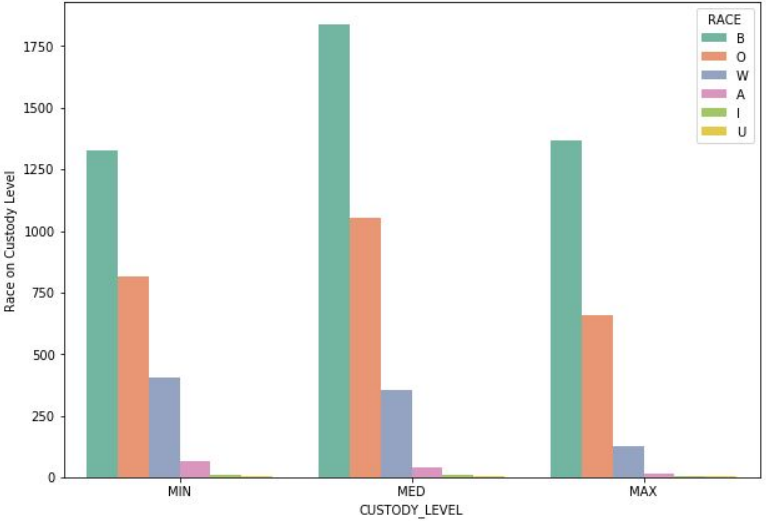
1. 160.15 is 'Robbery in the first degree'.
2. 160.10 is 'Robbery in the second degree'.
3. 110-125.25, if treated as a range, is 'Attempt to commit a crime' to 'Murder in the first degree'. No singular offense was found for this charge.
4. 265.03 is 'Criminal possession of a weapon in the second degree'.
5. 125.25 is 'Murder in the second degree'.
6. 105.15 is 'Conspiracy in the second degree'.
7. 140.25 is 'Burglary in the second degree'.
8. 120.00 is 'Assault in the third degree'.
9. 110-160.15, if treated as a range, is 'Attempt to commit a crime' to 'Robbery in the first degree'. No singular offense was found for this charge.
10. 110-160.10, if treated as a range, is 'Attempt to commit a crime' to 'Robbery in the second degree'. No singular offense was found for this charge.

Meeting the Goals

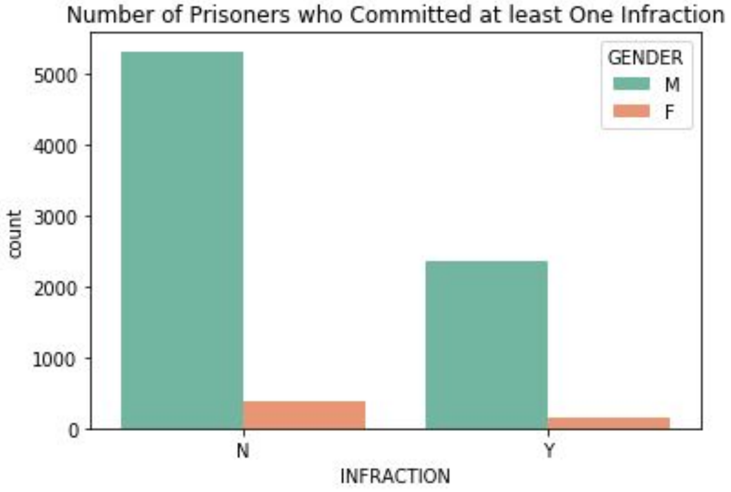
Neither race nor gender have overly large influence on custody level

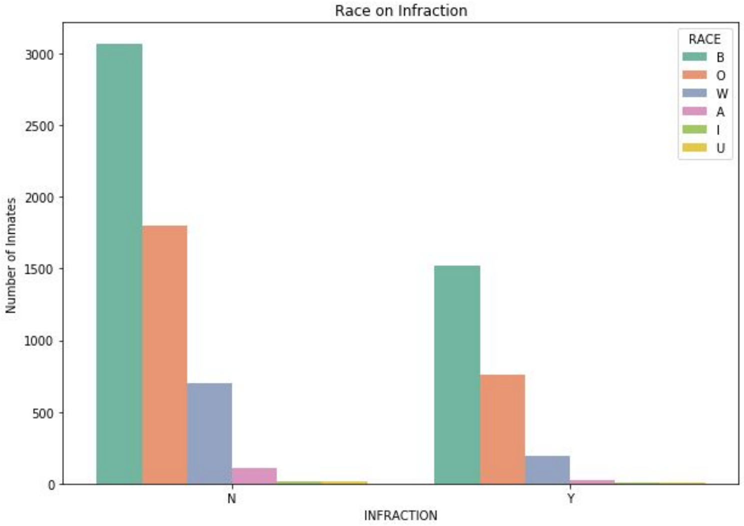


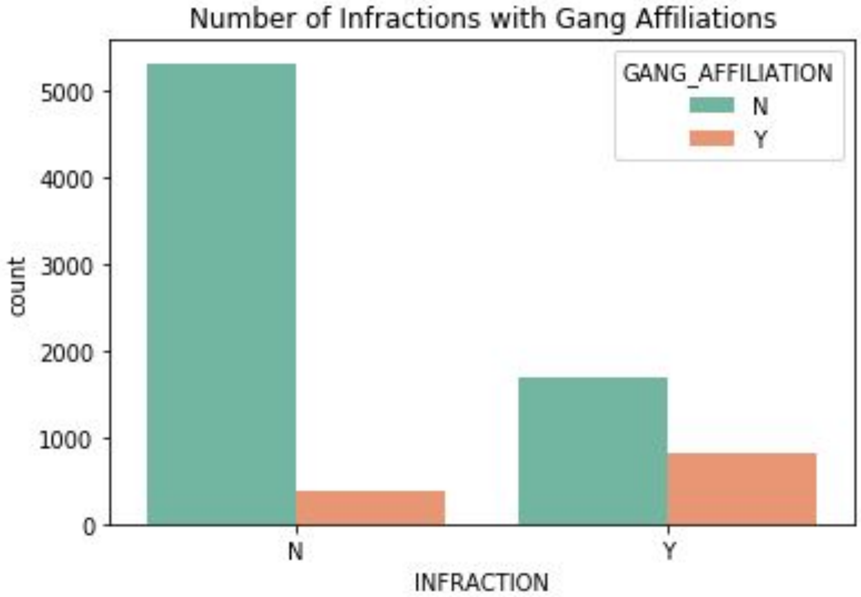
```
Text(0,0.5,'Race on Custody Level')
```



There is a pattern in infractions amongst certain groups of inmates.





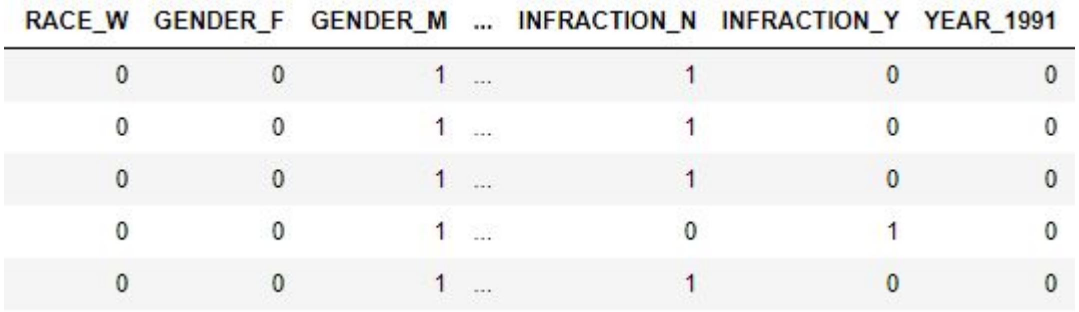


There are no patterns between race and gender and the number of infractions written. Of the people who do commit infractions, a majority of them are not in gangs. Although more gang affiliated inmates commit infractions than don't.

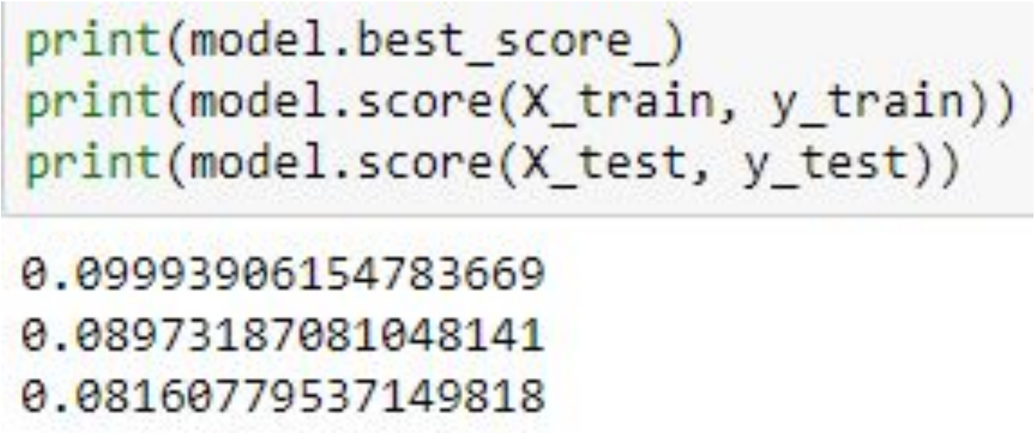
Machine Learning (also meets the third goal)

Preprocessing

- Dropped the AGE feature from the dataframe.
- Dropped the MONTH, DAY, and HOUR features from the dataframe.
- Dropped INMATE_ID.
- Applied one-hot encoder to convert categoric labels into numeric.



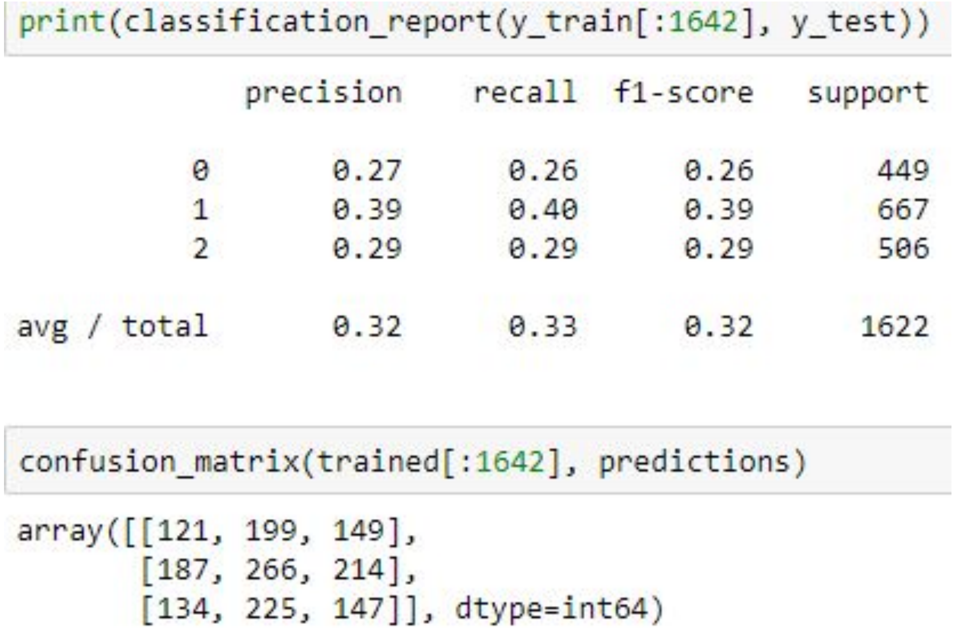
RACE_W	GENDER_F	GENDER_M	...	INFRACTION_N	INFRACTION_Y	YEAR_1991
0	0	1	...	1	0	0
0	0	1	...	1	0	0
0	0	1	...	1	0	0
0	0	1	...	0	1	0
0	0	1	...	1	0	0



```
print(model.best_score_)
print(model.score(X_train, y_train))
print(model.score(X_test, y_test))
```

```
0.09993906154783669
0.08973187081048141
0.08160779537149818
```

The scores provided are of the metric R^2. We can see that the accuracy of the test set was only 0.0081 less than the accuracy of the train set. However, the best score is more than 0.01 better (1.02%).



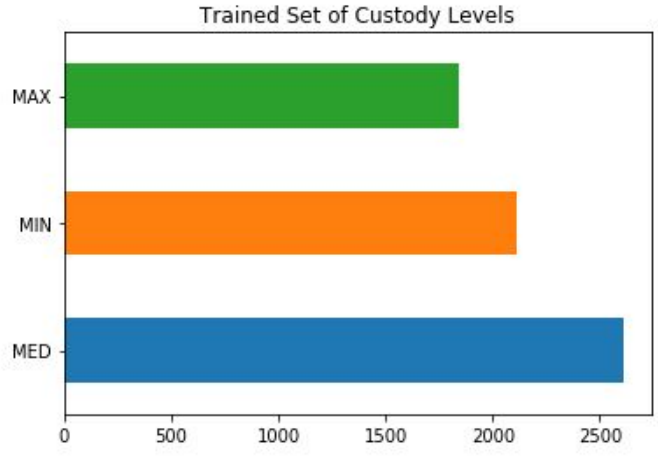
```
print(classification_report(y_train[:1642], y_test))
```

```
             precision    recall  f1-score   support

          0       0.27      0.26      0.26       449
          1       0.39      0.40      0.39       667
          2       0.29      0.29      0.29       506

avg / total       0.32      0.33      0.32      1622
```

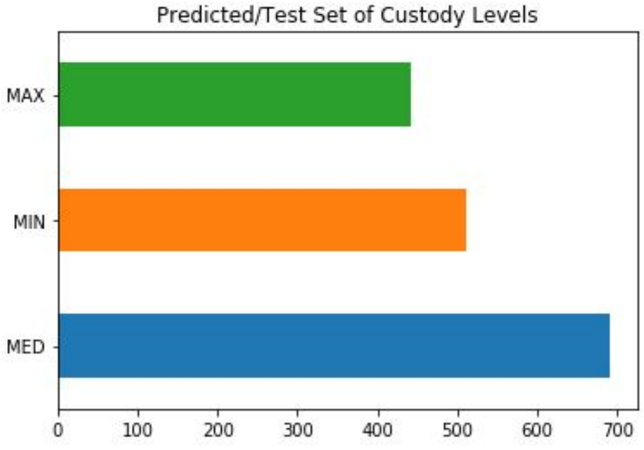
```
confusion_matrix(trained[:1642], predictions)
```

```
array([[121, 199, 149],
       [187, 266, 214],
       [134, 225, 147]], dtype=int64)
```

The precision of 0.32 shows that the classifier had more false positives in its predictions than true positives. The recall of 0.33 shows that the classifier was only able to find one-third of all the positive samples. The f1-score of 0.32 is the weighted average of the precision and recall where its best value is at 1 and its worst value is 0. Also note that the confusion matrix is not diagonal.

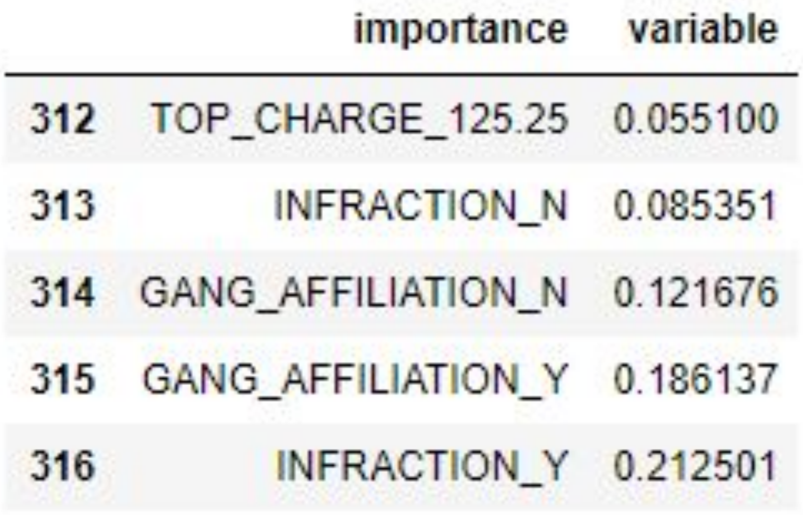


Trained Set of Custody Levels



Predicted/Test Set of Custody Levels

Graphically the classifier appears to do better than the classification report or the confusion matrix describe. While the outcome of the classifier appears to put the same proportions of inmates in respective custody levels, it does not appear as if the classifier made excellent use of the dataframe’s categorical variables to arrive at its conclusions.



	importance	variable
312	TOP_CHARGE_125.25	0.055100
313	INFRACTION_N	0.085351
314	GANG_AFFILIATION_N	0.121676
315	GANG_AFFILIATION_Y	0.186137
316	INFRACTION_Y	0.212501

As we can see the features that had the most influence on the custody level of the inmate was whether or not they were gang affiliated and had committed any infractions. The top charge value 125.25 is ‘Murder in the first degree’. Even that was taken into less of consideration by the model than gang affiliation and infractive behaviour.

Suggestions

1. Future datasets should have also have the discharge dates in them so as to create a more robust analysis. An analysis in the length of time spent incarcerated against, race, gender, age, custody level, infraction, top charge, mental observation, and gang affiliation would be more enlightening into patterns in NY daily inmates in custody.
2. The number of missing top charges in the dataset is going down by the year. Continuing that, the missing top charges should not be a problem by the end of 2019.
3. There are such a disproportionate number of men towards women in the dataset. 14.45 times as many men than there are women incarcerated. NY should look into creating and implementing social programs that aim to reduce the amount of criminal behaviour found 14x more often in men than women.