Baltimore Crime Data Classification

By: Nicholas Ogbonna

Background

The Baltimore Police Agency launched a massive overhaul to its new Records Management Systems back in May 2020. The improvement will enable the department to migrate from a paper-based system to a completely digital reporting environment. As a consequence of this major shift, we had significant difficulties in appropriately transferring data from the new records system to the previous Open Data Baltimore system. The "Arrests" dataset is one of many open datasets made publicly accessible by Baltimore's police department on the city's Open Data website. This information is provided to us in order to foster more openness and data exchange between the local administration and its residents. This dataset contains arrest records for offenses such as assault, theft, and property damage in the City of Baltimore. [1]

Business Question: Can I create a model that can predict when an arrest will occur based on data openly available?

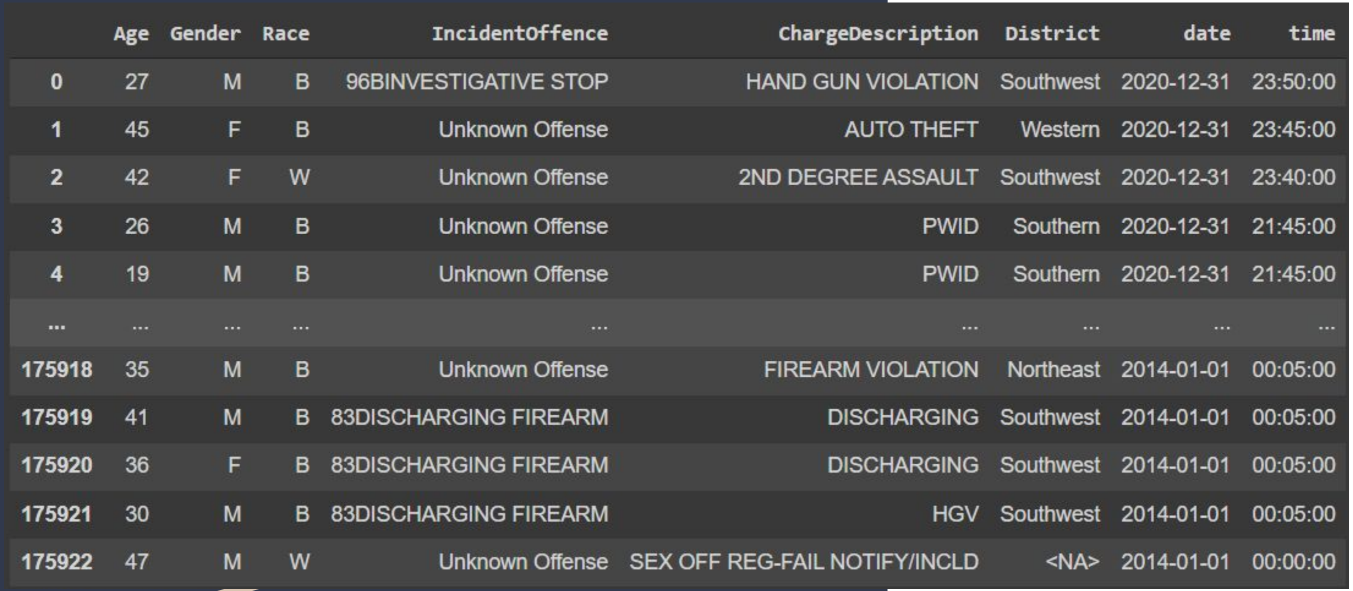
I want to use their database create a model that predicts which district a crime occurs based on various details related to the arrest of the perpetrator such as what he or she was charged with, their gender, etc. I will consider the model to be a success if it can predict the District a crime occurred at least 80% of the time. My hope with this model is that the department can then use this model to efficiently spread their resources tackling the more likely arrests that would be made in a certain district as well as modify their policing efforts to decrease bias in said policing efforts in certain districts (if any).

EDA Results

First thing I notice is that there are some features that will not contribute to my classification problem.

- 'X', 'Y', 'RowID', and 'ArrestNumber': are all similar indexing variables that will provide nothing to the classification
- 'ArrestLocation', 'IncidentLocation', 'Neighborhood': This wouldn't fit my classification problem as my model shouldn't need location information to determine the district the crime happened
- 'Latitude', 'Longitude', and 'GeoLocation' all provide the same information. This wouldn't fit my classification problem as my model shouldn't need location information to determine the district the crime happened
- 'Shape': Shape is a useless feature column that provides no information and is mostly filled with NA values
- Charge, post: Charge and Charge Description hold what is basically the same information. Post is a 3 digit code relating to someone's charge. Thus when I throw them into a OHE they will provide redundancy which is bad.

My Modeling Data



	Age	Gender	Race	IncidentOffence	ChargeDescription	District	date	time
0	27	M	B	96BINVESTIGATIVE STOP	HAND GUN VIOLATION	Southwest	2020-12-31	23:50:00
1	45	F	B	Unknown Offense	AUTO THEFT	Western	2020-12-31	23:45:00
2	42	F	W	Unknown Offense	2ND DEGREE ASSAULT	Southwest	2020-12-31	23:40:00
3	26	M	B	Unknown Offense	PWID	Southern	2020-12-31	21:45:00
4	19	M	B	Unknown Offense	PWID	Southern	2020-12-31	21:45:00
...	...	...	...	...	...	...	...	...
175918	35	M	B	Unknown Offense	FIREARM VIOLATION	Northeast	2014-01-01	00:05:00
175919	41	M	B	83DISCHARGING FIREARM	DISCHARGING	Southwest	2014-01-01	00:05:00
175920	36	F	B	83DISCHARGING FIREARM	DISCHARGING	Southwest	2014-01-01	00:05:00
175921	30	M	B	83DISCHARGING FIREARM	HGV	Southwest	2014-01-01	00:05:00
175922	47	M	W	Unknown Offense	SEX OFF REG-FAIL NOTIFY/INCLD	<NA>	2014-01-01	00:00:00

These are the features that I will use to train my model.

Modeling Results

- None of the methodologies I used were able to produce an accuracy over 25%. This model would be unsatisfactory to use in production if we wanted to try and predict where an arrest would occur.

My Project for 604

The methods I used did not produce an accuracy rate of over 25%. This did not come anywhere near my 80% threshold leading me to believe the following:

- The seven features alone are insufficient to adequately identify the District.

- Insufficient rows were utilized to train the data.

- the methodologies were insufficient to predict the District with just a few columns of data.

My Project for 604

I want to see weather SPARK's clustering and various other tools can help improve my prediction rate & see whether there is any bias in policing efforts.

If I have time, I would also like to perform:

- Semi-Supervised Learning
- SVM modeling
- ADABOOST

On whatever SPARK produces to see if my prediction rate increases

References

[1] “Baltimore Police Department.” Crime Stats | Baltimore Police Department, https://www.baltimorepolice.org/crime-stats.