Final Project Check-in

By Nicholas Ogbonna

Problem statement and Data Source

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]

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. 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).

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

Initial data exploration-- types, distribution, values, etc.

X	Float64
Y	Float64
RowID	Int64
ArrestNumber	Int64
Age	Int64
Gender	string
Race	string
ArrestDateTime	string
ArrestLocation	string
IncidentOffence	string

IncidentLocation	string
Charge	string
ChargeDescription	string
District	string
Post	string
Neighborhood	string
Latitude	Float64
Longitude	Float64
GeoLocation	string
Shape	Int64

dtype: object

175923 rows × 8 columns

Unnecessary features

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.

Current workspace

After dropping the unnecessary features and modifying the timestamp to a more useful date & time, I have the following features

Age	Int64
Gender	string
Race	string
IncidentOffence	string
ChargeDescription	string
District	string
date	object
time	object

Basic Info

	count	unique	top	freq	mean	\
Age	175923.0	NaN	NaN	NaN	32.949262	
Gender	175923	3	M	142091	NaN	
Race	175923	6	B	144592	NaN	
IncidentOffence	175923	172	Unknown Offense	91417	NaN	
ChargeDescription	175923	7028	FAILURE TO APPEAR	27580	NaN	
District	89891	9	Southern	11771	NaN	
date	175923	2866	2014-01-09	177	NaN	
time	175923	1440	11:00:00	2834	NaN	

	std	min	25%	50%	75%	max
Age	11.441689	0.0	24.0	30.0	40.0	100.0
Gender	NaN	NaN	NaN	NaN	NaN	NaN
Race	NaN	NaN	NaN	NaN	NaN	NaN
IncidentOffence	NaN	NaN	NaN	NaN	NaN	NaN
ChargeDescription	NaN	NaN	NaN	NaN	NaN	NaN
District	NaN	NaN	NaN	NaN	NaN	NaN
date	NaN	NaN	NaN	NaN	NaN	NaN
time	NaN	NaN	NaN	NaN	NaN	NaN

Date and Time information

Arrest Data is logged from: 00:00:00 to 23:59:00

Arrest Data Is collected All day from:

2014-01-01

To

2021-11-05

Interesting tidbits from the 5 number summary

Just from the 5 number summary alone, we got a wealth of information:

The most common way to get arrested is to not show up to your appointed court date. Whether it's a major or minor offence, one should always try to make their court date.

Most arrests seem to be made around 11am.

The most amount of arrests made on a single day was on January 9th 2014

In the 7 years this data spans, the black males are the most likely to be arrested. There are many articles backing up policing bias that contributed to this trend. I will leave some articles below.

Your proposed solution

- I’ve recently learned that the model i’ve created is actually twice as effective as naive, but I **know** I can get better than 23% logistic regression gave me.
- I’ve successfully gotten and ensemble to run and while it showed favorable results. It's not as good as pure logistic regression
- Currently, my solution is to find a day (may 4th from 7pm to 10pm looks good) where I can leave my laptop and have all cores focus on running SVM to see if it’d even do any better than ensemble and ADABoost.