Analyzing Arrest Data from 2010 to Present in Los Angeles

Christopher Cerritos, Tam Do, Mike Avelar
Department of Information Systems
California State University, Los Angeles

Abstract: The Los Angeles Police Department[1] has transcribed original arrest reports typed on paper into a public digital dataset. The Los Angeles Open Data site shares city data with the public in a mass repository to provide useful information to individuals, companies and organizations to further improve the well-being of the city. With that in mind, publicly published data allows transparency for the public sector to engage common citizens to delve into research, advancements, develop and innovation for the city. With this dataset, various types of arrest cases within different area zones were counted and recorded, as well as individual biometrics, including age and gender. Using Microsoft Azure Machine Learning (ML) Studio, the dataset was cleaned and used to train Bayesian Regression models in various configurations.

1. Introduction

The goal of this project was to accurately analyze one area possible alternative to reduce arrest rates given the number of arrests made within the specified area. From the data reported by the Los Angeles Police Department, original arrests transcribed on paper have been electronically filed from 2010 to present. Every record in the dataset represents a real arrest, referenced by its location, time, charge and limited demographics pertaining to the individual.

2. Related Work and Background

This data provides significant information of an individual's arrest and their committed crime. As such, it helps determine possible solutions for these high-crime areas.

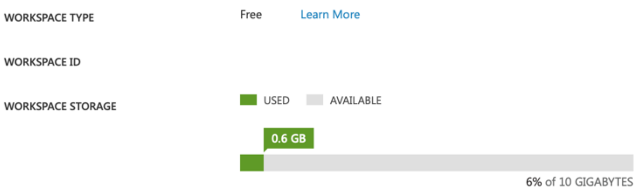
3. Platforms Used

This project was conducted using one cloud solution platform and another locally installed software. The data cleaning, trimming, transformation and testing of the model was performed on a free tier version of Microsoft Azure Machine Learning Studio. Some of that data was also imported to Elasticsearch and Kibana to help provide visualizations that provided a better understanding of raw data.

3.1 Microsoft Azure Machine Learning Studio

Microsoft Azure Machine Learning Studio is a managed cloud platform owned by Microsoft, therefore not much information is publicly available regarding its specifications and components behind the application. Although, they do offer different service levels which can further enhance an end-user's experience when using the tool. Our teams testing was conducted on a free tier version, provided by our school.

In the free tier service, the service is limited compared to those paid tiers. Although, just like Amazon AWS, you pay only for what you use. Storage space used for saving the datasets used in this project is limited to 10GB. Our experiments in this project only ran at low priority on a single processor core. This resulted in longer wait times.



WORKSPACE TYPE Free Learn More

WORKSPACE ID

WORKSPACE STORAGE 🟩 USED ▢ AVAILABLE

0.6 GB

6% of 10 GIGABYTES

Figure 1. Specifications of Microsoft Azure Machine Learning Studio free-tier.

3.2 Elasticsearch and Kibana Service

Like Azure ML, the Elasticsearch and Kibana service is also a managed cloud service. However,

[1] Dataset provided by the Los Angeles Police Department on a weekly basis.

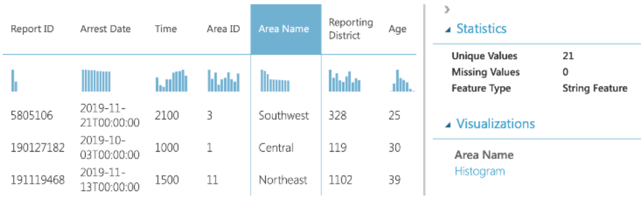
the team preferred to install a local instance of Kibana to connect to a remote cluster of Elasticsearch.

4. Data Preparation

The data used for this project was a 222.8MB dataset downloaded in a CSV file obtained from the Los Angeles Open Data [1] repository, downloaded on December 14, 2019. There are several export methods of the data but our best interest was contained in a CSV file. We can agree that the data on this file is accurate as it is reported on a weekly basis from a credible source, the Los Angeles Police Department.

4.1 Area

For obtaining information relating to the area of these arrests, we referenced to the raw data export and reviewed the Area ID and Area Name columns. Each Area ID represented a unique value for its Area Name.



Report ID	Arrest Date	Time	Area ID	Area Name	Reporting District	Age
5805106	2019-11-21T00:00:00	2100	3	Southwest	328	25
190127182	2019-10-03T00:00:00	1000	1	Central	119	30
191119468	2019-11-13T00:00:00	1500	11	Northeast	1102	39

Statistics

Unique Values	21
Missing Values	0
Feature Type	String Feature

Visualizations

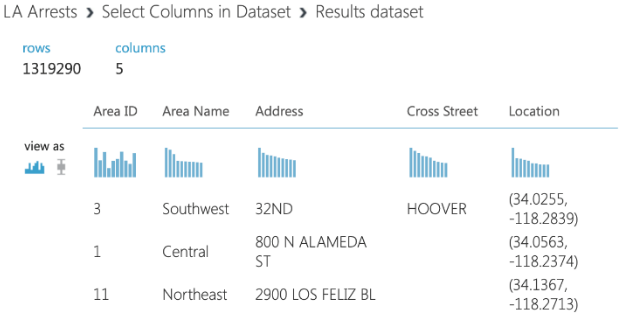
Area Name
Histogram

Figure 2. The first few columns of the “Arrest_Data_from_2010_to_Present.csv”

Area ID	Area Name
1	Central
2	Rampart
3	Southwest
4	Hollenbeck
5	Harbor
6	Hollywood
7	Wilshire
8	West LA
9	Van Nuys
10	West Valley
11	Northeast
12	77th Street
13	Newton
14	Pacific
15	North Hollywood
16	Foothill
17	Devonshire
18	Southeast
19	Mission
20	Olympic
21	Topanga

Table 1. Detailed reference of Area ID and Area Name columns.

For each Area ID, the ID denotes the Area Name of the location where the arrest was made. A total of 21 unique values were found as we analyzed the areas of interest.



LA Arrests › Select Columns in Dataset › Results dataset

rows	columns
1319290	5

Area ID	Area Name	Address	Cross Street	Location
3	Southwest	32ND	HOOVER	(34.0255, -118.2839)
1	Central	800 N ALAMEDA ST		(34.0563, -118.2374)
11	Northeast	2900 LOS FELIZ BL		(34.1367, -118.2713)

Figure 3. Columns related to the area of the arrest.

5. Results

For this project, several different configurations f prediction models were used. The 21 unique areas identified in this dataset helped analyze areas in need of resources to reduce crime arrests. We took this data very seriously to look at correlations between arrest rates and charge codes within district areas.

As a group, it was in our interest to look at how the Van Nuys area compared to other areas within the LAPD community police stations. The 21 geographic areas or patrol divisions are refenced by unique IDs and assigned to a nearby community police station.

We struggled to get to a result with the use of Azure ML Studio. However, we think that we were going in the right direction. We put together a lab manual[2] that demonstrates our efforts to get a result for our objective.

Apart from uploading our dataset to Azure ML Studio, it was in our interest to train the raw

[2] Lab manual can be found in our GitHub repository.

dataset. With this, we were looking at Area ID, Area Name, Age and Sex Code columns.

6. Visualizations

The dataset used in the models as well as other interesting data were imported into Elasticsearch and visualized in manually created charts. Due to the limitations of the free tier software, we had to cut down our data to 100MB as the visualizer feature was expiremental.

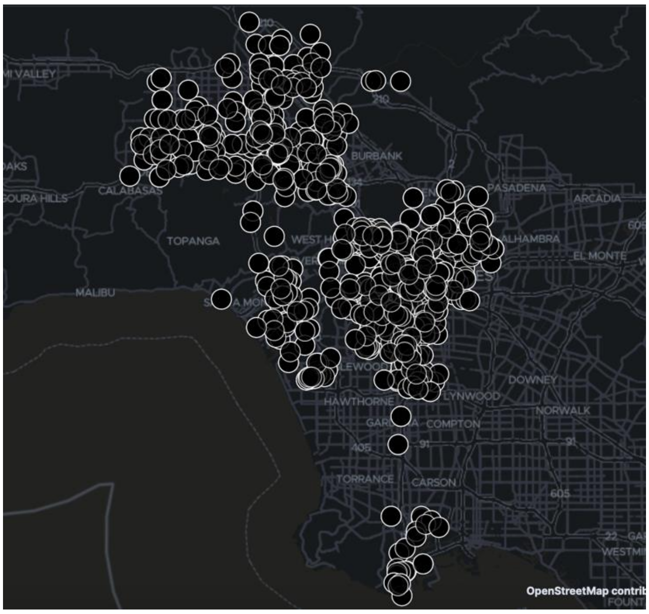
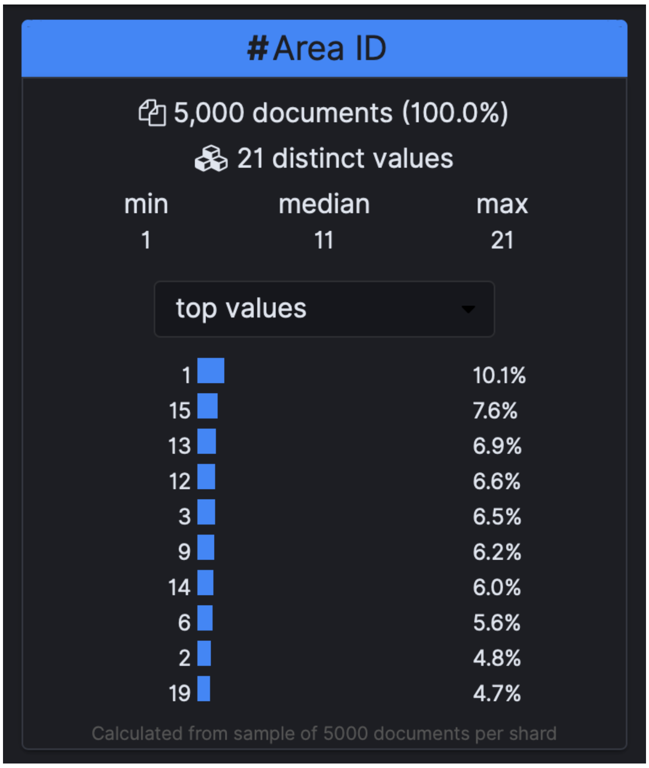


Figure 4. A map visualizing arrest location.

From the GeoJSON vector file of our modified dataset, we imported the file to index in Elasticsearch. As Figure 4 suggests, all the arrests made that were recorded in the raw data are noted here. Although the results are zoomed to layer, we can visualize that some areas have more activity than others.



#Area ID

5,000 documents (100.0%)

21 distinct values

min	median	max
1	11	21

top values

1	10.1%
15	7.6%
13	6.9%
12	6.6%
3	6.5%
9	6.2%
14	6.0%
6	5.6%
2	4.8%
19	4.7%

Calculated from sample of 5000 documents per shard

Figure 5. With randomly selected rows, we can see how the Van Nuys (9) area compares to other neighboring areas.

The work found in in Elasticsearch was done for data visualization purposes. We wanted to extract the data from our downloaded dataset to visualize the differences amongst the age groups, total arrests within areas and top charges for these arrests. More visuals can be found in our PowerPoint presentation[3].

7. Conclusion

This project allowed us to test several regression models within Azure ML Studio. Our objective was to define areas with most arrest cases to possibly suggest methods of improvements in the community. Our initial dataset was not based on Arrest Data from 2010 to Present in Los Angeles, rather we planned to look at Auto Collisions in the City of New York. Though we had difficulties we decided it'd be best to look at data within our own county.

With the use of Azure ML Studio and past in-class lab exercises, we put together a diagram with different modules for predictive analysis in

[3] PowerPoint presentation can be found in our GitHub repository.

the Van Nuys area. We built an experiment involving Bayesian Linear Regression, Two-Class Decision Forest and Permutation Feature Importance[4]. Through these modules, we tried to manipulate our raw dataset.

Expanding our analysis to Elasticsearch, with the use of Kibana we were able to visually prepare graphs to better look at the raw data. Due to the limitations of the free tier of the software, we were not able to include the full dataset. However, with the 590,000 records that were imported, it was enough to get an idea of the data and its definition.

Based on the charge codes and areas of most arrests, the county of Los Angeles must allocate resources in the area to reduce the statistics in the area. With the current data, it suggests that the number of arrests and infractions generalize the area as unsafe for innocent people.

References

[1] Arrest Data from 2010 to Present, *Los Angeles Open Data*. https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4

[2] https://github.com/ccerritoss/cis-3200

[4] Created expirements can be found in Tam Do’s Azure ML Studio.