Safety Classification

Jean Chang, CSc 59927, Spring 2018

***Abstract*—The classification of hazardous areas with high motor vehicle accidents allows governmental agencies responsible for improving road safety to properly allocate the areas that are prone to accidents. In this paper, hazardous areas are identified by factoring in the number of accidents, injuries, deaths, and number of vehicles. The results show that the top vehicle type is passenger vehicle and top street complaint is street light out.**

***Index Terms*—Road safety; Road accidents; Big Data; Apache Spark; Data Frame**

I. Motivation

According to the NYS DMV, between 2011 and 2013 there were 232,267 motor vehicle accidents in New York City, resulting in 833 deaths. In 2014, the Vision Zero Action Plan was announced and its goal is to eliminate all traffic deaths and injuries in NYC. When the city debated about the causes of these accidents and how to prevent more traffic deaths and injuries, the DOT released a dataset of all NYC traffic crashes from 2008 to present.

II. Project Objectives

The main objective is to identify hazardous areas around NYC that are prone to motor vehicle accidents by factoring in the number of accidents, injuries, deaths, and number of vehicles involved and profile these hazardous areas. The plan is to develop a classification algorithm for classifying the areas, analyze 311 complaints related to street/traffic conditions and type of vehicle, then plot areas labeled as hazardous with a profile (most common vehicle type, most common 311 traffic/street conditions complaints). My hypotheses are the most common vehicle type are commercial or taxi, and the most common 311 complaints around areas with most accidents will be potholes.

III. Data Sources

For this project, three datasets are used and they are all provided from NYC OpenData. The datasets are NYPD Motor Vehicle Collisions, Traffic Volume Counts, and 311 Service Request. The size of the 311 Service Requests is 4.69 GB, NYPD Motor Vehicle Collisions is 273.9 MB, and the Traffic Volume Counts is 929 KB. The data is analyzed from 2012 to 2013 because the data from traffic volume from DOT is only available from 2011 to 2013.

IV. Big Data Challenges

The biggest challenge is the data preparation stage of the data analytics lifecycle. This is because of the number of dataset used as there are 146 fields and over 10,000,000 rows over the three datasets. Only the relevant fields will require extensive normalization. Other challenges included variety among values in common field. For example, the NYPD motor vehicle collisions and vehicle classification count datasets both provide a field for the type of vehicle but using different terms. Data cleansing will also be needed because many of the datasets have missing values.

V. Methodologies

A classification algorithm supported by Spark was developed that will label a given location as ‘hazardous’, ‘ok’, or ‘safe’ based on the factors listed above. The average and standard deviation will be computed in four feature categories for each area via zip code.

1. Accident/1000 vehicles
2. Deaths/1000 accidents
3. Injuries/1000 accidents
4. Vehicles involved/accidents

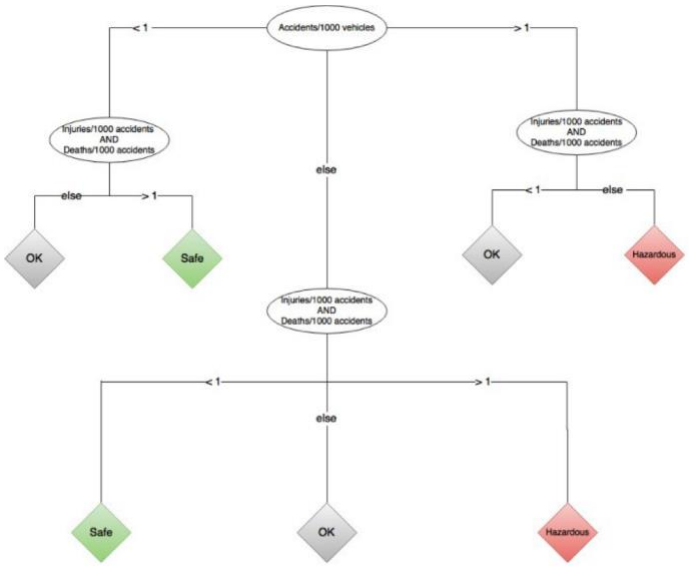

Fig. 1. Decision tree of the classification algorithm.

In the decision tree <1 means one standard deviation below

mean and >1 means one standard deviation above mean. This is how the algorithm classify the features.

The idea is that those areas where multiple feature have values greater than one standard deviation from the mean are the most hazardous.

VI. Source Code

For the police report file, the NYPD Motor Vehicle Collisions dataset is read into a RDD. Then parses each accident record and extracts relevant fields like zip code and the number of persons injured for records between 2012-2013, with zip code as the key. Next, takes data for each zip code and use a dictionary to aggregate various statistics for that zip code. These include the total number of accidents, total number of vehicles involved in accidents, and total number of persons injured. Each zip code has a dictionary containing those relevant statistics. After that, all the dictionaries are merged for each zip code into one dictionary containing the metadata for the entire set and is saved. Finally, run two heapq.nlargest() operations to determine the top five accident factors and top five vehicle types involved in accidents for each zip code. If a zip code that had fewer than five unique accident factors or vehicle types involved in accidents, the top three are computed. If there are fewer than three types of factors, the top one is computed. The results are saved in the raw results folder

For three_one_one file, the 311 dataset is read into an RDD. Then the dataset is filtered for street related complaints in 2012-2013, and any record that do not have a zip code or city associated with them are disregarded. Next, aggregates the street complaint related data for each zip code in a dictionary. Keys in this dictionary include the total number of complaints, and a count for each complaint type and description. Then all the dictionaries are combine into one that contain metadata for the dataset. Finally, the dictionary associated with each zip code, and computes the top 3 complaints types and top 5 complain descriptions for that zip code. If there are top 3 complaint types, then the top one is computed. If there are fewer than five complaint descriptions, then the top three are computed. If there are fewer than three, then the top one is computed. The results are saved in the raw results folder.

For vehicle volume count file, the zip_veh_count file is read into an RDD. The map_mutiple_zip method find these records that have two zip codes associated with them and splits the vehicle count and samples between the two zip codes in the record. It yielded a record with a single zip code as the key, and a vehicle count and number of samples. Then a dictionary is produced for each zip code, which has aggregations for the vehicle count and number of samples. The final normalized result is computed with a key equal to the zip code, and a lone value equal to the normalized traffic volume for 2012-2013. The results are saved in the raw results folder.

For main file, a single RDD called joined_rdd for the three pervious files, and a DataFrame for the RDD are created. This makes it easier to compute the mean and standard deviation for each feature by performing a single call on the DataFrame. The top 10 and bottom 10 zip codes for each feature are also computed using the DataFrame. The results are saved in the raw results folder.

All code is available at https://github.com/jchang96/Safety-Classification.

VII. Results

The results are classified by using the decision tree described above.

City	Zip Code	Top 3 Factors	Top 3 Vehicle Type	Top 3 Street Complaints
Bronx	10468	Driver Inattention/ Distraction Failure to Yield Right of Way Fatigued/ Drowsy	Passenger Vehicle Sport Utility/ Station Wagon Taxi	Pothole Street light out Failed street repair
Flushing	11352	Driver Inattention/ Distraction Failure to Yield Right of Way Fatigued/ Drowsy	Passenger Vehicle Sport Utility/ Station Wagon Other	Street light out Pothole Rouge, pitted, or cracked roads
Brooklyn	11231	Driver Inattention/ Distraction Oversized Vehicle Backing Unsafely	Passenger Vehicle Sport Utility/ Station Wagon Other	Street light out Pothole Failed street repair

Table. 1. Top 3 most hazardous zip code with factors, vehicle types, and street complaints.

Table 1 shows that the top three factors are driver inattention/distraction, failure to yield right of way, and fatigued/drowsy, top three vehicles types are passenger vehicle, sport utility/station wagon, and other, and top three street complaints are street light out, pothole, and failed street repair.

Label	Number of Classifications	Normal Distribution Expected
Hazardous	18 (12.5%)	16%
Ok	117 (81.25%)	68%
Safe	7 (9.25%)	16%

Table. 2. Percentage of classification for each label with normal distribution expected.

Table 2 shows the percentage of classifications for each

label, ‘hazardous’, ‘ok’, or ‘safe’ with the normal distribution expected.

VIII. Deliverable

The type of data visualization created is a 2D choropleth map of the five boroughs partitioned by the zip code. The map is created with Carto.

The map shows that areas in south-east Brooklyn, patches in the Bronx, and patches in Queens are labeled as hazardous, meaning those areas are more prone to motor vehicle accidents.

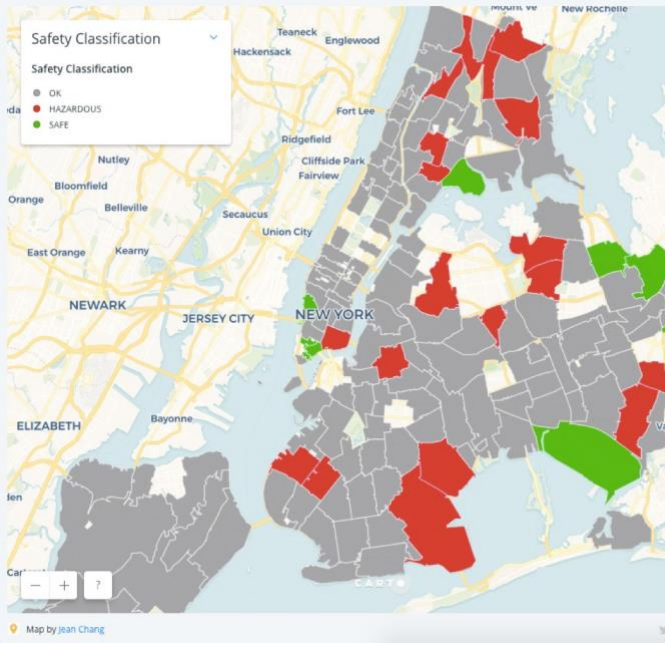

Fig. 2. 2D choropleth map of areas in NYC classified as hazardous, ok, or safe.

IX. Conclusion

The biggest challenge was changing the classification algorithm. The initial method for classification of zip codes was to use Spark’s Machine Learning library decision tree classifier to classify areas as ‘safe’, ‘ok’, or ‘hazardous’. However, because there are only 144 zip codes compared to the large dataset of 311 complaints, a different method was used. And that methods id to compute the average and standard deviation in three feature categories for each zip code.