Investigating Motor Vehicle Crashes in New York City

Casey Delaney, San Jose State University[1],
Manasa Bobba, San Jose State University[1], and
Sudheendra Katikar, San Jose State University[1]

[1]

1. Abstract

Ongoing studies have anticipated that in 2030, car crashes will be the fifth driving reason for death around the world. Road accidents and its safety have been a major concern around the world and have been a primary concern for automotive manufactures. Road safety concerns sparked the creation of the NHTSA (National Highway Traffic Safety Administration) as well as the IIHS (Insurance Institute for Highway Safety) among others. Road accidents and reckless driving occur in every part of the world. Because of this, many pedestrians are affected too. This paper aims to analyze road accidents in one of the most populous cities in the United States, New York. The data we have used is especially informative because it is generated from a comprehensive police report template. It consists of information from all police reported motor vehicle collisions in NYC over the period of 2012-2021 and has over 1.8 million records.

2. Introduction

This paper has the purpose of determining contributing factors that cause higher instances of accidents in automobiles. In a span of 9 years, over 1.8 million accidents were documented in New York. By analyzing these traffic records, billions of dollars worth of damage may be recovered in future years. A 2014 report showed that 871 billion dollars per year is being wasted due to traffic collisions in the United States [Lowy 2014].

2.1 Domain introduction

2.1.1 Machine learning
Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of Artificial intelligence. Machine learning algorithms build a mathematical model based on sample data known as training data. Machine learning algorithms are widely used in various applications such as email filtering, computer vision. It is mainly focused on computational statistics which focuses mainly on predicting.

2.1.2 Data mining
Data mining is a branch which involves looking for hidden, valid, and potentially useful patterns in huge data sets. It is all about discovering previously unknown relations among data. Data mining involves anomaly detection, association rule mining, clustering, classification, regression and more.

2.2 Data introduction
The dataset used is NYPD's Motor Vehicle Collisions, which is provided by the NYC Open Data repository. The dataset consists of information about all collisions in NYC. It contains information about date, location, cause of the accident, number, and type of injured, and killed road users. The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. There are 1.8 million rows and 29 columns.

2.3 Methods
Three methods of analyzing the data will be used. Each method has a specific purpose and will show features in the data that other methods are unable to reproduce. The first method is association analysis with the A-Priori algorithm. This algorithm allows common relationships in data item sets to be found. For example, are there any associations with certain streets and types of accidents? The algorithm is especially useful in finding matching items between different columns. The K-Means clustering algorithm will be used to find a visual relation between large groups of data. This algorithm will be used to determine if a certain longitude/Latitude pair have higher incidents of accidents.

2.4 Initial Data Impressions
Performing initial data analysis has provided some inconclusive results. Figure 1 shows the number of accidents as a function of the month it occurs in. New York can contain harsh winter conditions during the months of November through April. One might hypothesize that this would cause the number of accidents per month to increase. The data shows the opposite with the number of accidents per month peaking in June at 160,000. This could be because more people choose to stay indoors during the winter months.

Figure 2 shows the number of accidents per year. From the years 2012 through 2018, there has been a steady increase in accidents. This can be due to a variety of factors such as

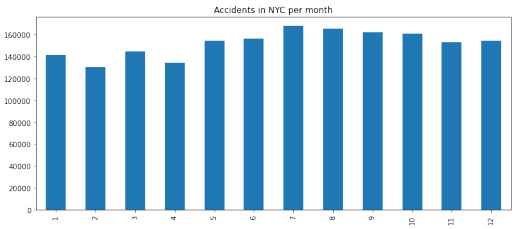

Fig. 1. Number of accidents as a function of month in NYC

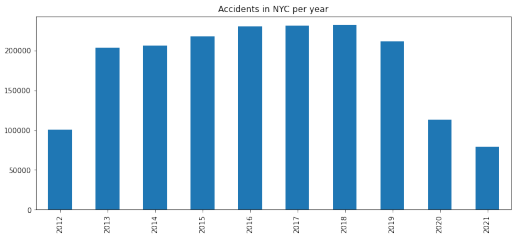

Fig. 2. Number of Cyclist Injuries as a function of year in NYC

increasing population putting more vehicles on the road. In 2019, there is a small decline in the number of accidents and 2020 shows a much sharper decline. This can be a direct result from the Covid-19 pandemic creating a work-at-home situation for many citizens. The data for 2021 is incomplete since the year is not over at the time of this writing (data downloaded in October, 2021) and will therefore not be considered for this impression.

Interesting statistics begin to appear when the number of injuries are viewed. Figure 3 shows the number of cyclist injuries as a function of year. In the past 3 years, there has been a dramatic increase of cyclist injuries. The year 2021 is still incomplete at the time of this writing, yet there are more cyclist injuries in 2021 than in any of the previous years.

Finding relationships for the causes in these injuries has the potential to save many lives each year. Although the total number of deaths in New York City may seem small, this is only one city among many throughout the United State and the rest of the world.

3. Methods

To analyze the data set, three methods were used. These methods consist of association analysis using the A-Priori

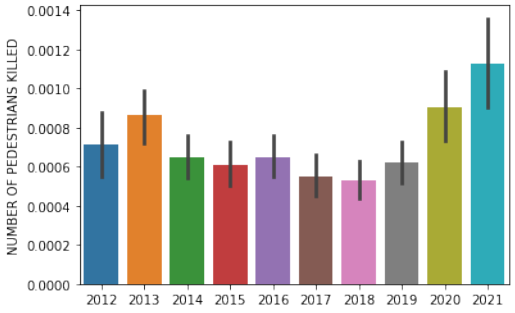

Fig. 3. Number of accidents as a function of year in NYC

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
6	(Failure to Yield Right-of-Way)	(QUEENS)	0.344999	0.254405	0.097481	0.282554	1.110649	0.009712	1.039236
7	(QUEENS)	(Failure to Yield Right-of-Way)	0.254405	0.344999	0.097481	0.383173	1.110649	0.009712	1.061888
4	(Driver Inattention/Distraction)	(QUEENS)	0.327053	0.254405	0.091251	0.279008	1.096711	0.008047	1.034125
5	(QUEENS)	(Driver Inattention/Distraction)	0.254405	0.327053	0.091251	0.358683	1.096711	0.008047	1.049320
0	(Failure to Yield Right-of-Way)	(BROOKLYN)	0.344999	0.312535	0.114145	0.330856	1.058620	0.006321	1.027380
1	(BROOKLYN)	(Failure to Yield Right-of-Way)	0.312535	0.344999	0.114145	0.365223	1.058620	0.006321	1.031860
3	(Driver Inattention/Distraction)	(MANHATTAN)	0.327053	0.260903	0.087971	0.268982	1.030964	0.002642	1.011051
2	(MANHATTAN)	(Driver Inattention/Distraction)	0.260903	0.327053	0.087971	0.337180	1.030964	0.002642	1.015278

Fig. 4. A-Priori Analysis

algorithm, K-means clustering and classification. Depending on the algorithm used, the size of the data set may be reduced to decrease the processing time. All computations were performed in a notebook hosted by Google Colab and certain actions were unable to be completed due to Google Colab limitations. The entire data set was used in applicable instances where the computing power did not exceed the limitations.

3.1 Association Analysis using A-Priori Algorithm

Association analysis will be used to find common relationships between different itemsets. The A-Priori algorithm is used to determine when correlations and co-occurrences exist. Generally the A-Priori algorithm performs slower than other methods on large data sets. The algorithm is relatively simple but is limited due to its slow performance. Increasing the speed is possible in this research by limiting the data to data deemed useful such as the primary vehicle involved in the car accident (Vehicle Type Code 1) and the number of pedestrians or passengers injured or killed. The majority of car crashes are limited to one or two vehicles rendering columns for vehicles 3, 4 and 5 almost useless. These columns, among others, were left out to improve computational efficiency.

Figure 4 shows results from association analysis with this method using the A-Priori algorithm. The results will be analyzed in the analysis section of this paper. Initial results show a strong correlation between certain types of accidents. A common theme of accident time is Failure to Yield Right-of-Way and also Driver inattention/distraction. These accident themes are particularly common in the boroughs of Queens, Brooklyn and Manhattan. Due to the very large number of possible combinations, the minimum support threshold was lowered to 0.05 to receive proper results.

3.2 Clustering with K-Means

The use of the K-Means algorithm will allow the data to be clustered based on the data set features. The algorithm has been run on columns of latitude and longitude and longitude data to determine if there are certain localities that are especially prone to accidents. To this point, data has been somewhat inconclusive. Figure 5 shows the results of the K-Means clustering. As seen in the figure, especially in the bottom part of the blue section, there are some points that are very dense in accidents. This occurs around the (-73.98, 40.75) area. Results were also analyzed to determine if there are clusters around a specific crash time and this far, results have been inconclusive. This could be because the number of clusters must be increased to account for the large variation within the city. The size of the dataset makes it difficult to use the K-Means algorithm due to the limitations of Google Colab so

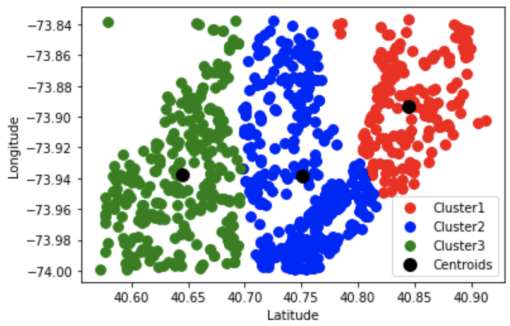

Fig. 5. K-Means Clustering as a function of longitude and latitude

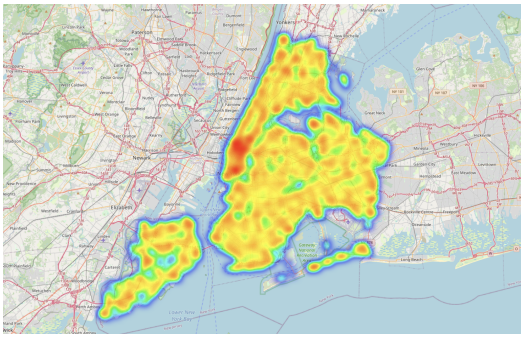

Fig. 6. Heat Map

the dataset was reduced in size.
Another method of performing a similar task was found by using a heat map. Using a Python Library, a map was created as shown in figure 6. This map shows accident densities in New York. Areas marked in red have a much higher likelihood of causing an accident. This provides a much more visually appealing graph to the viewer.

3.3 Classification
The categorical Naive Bayes classifier is suitable for classification with discrete features that are categorically distributed. It assumes that each feature has its own categorical distribution. In our use case, we attempted to build a classifier using the features 'CRASH TIME' and 'STREET NAME', to see if we could predict the cause of the accident, the class 'CONTRIBUTING FACTOR VEHICLE 1'. We encode the data using LabelEncoding, as we can't classify non-numerical features directly. We finally achieved an accuracy of 47 percent in predicting the cause of the accident. This result is visible in the code under the classification section.

4. Comparisons

The dataset we have taken is not the most used Kaggle dataset. There is only Explanatory data analysis (EDA) done on this dataset. No models were implemented in Kaggle. We have done almost every visualization possible on our data set.

5. Example Analysis

As seen in figure 4, a common theme in accidents consists of Failure to Yield Right-Of-Way in Queens. Why is this the case? This could be the result of a variety of factors. It is possible that the structure of the streets in Queens has created dangerous situations such as difficult turns where a driver may not yield to an oncoming driver due to the inability to see the road ahead. The purpose of this analysis is not to discover the reason why a certain condition occurs, but to identify this condition in order for local specialists to focus on a cause. There is also a large amount of driver inattention in Queens and Manhattan. A solution to this problem could be to run a short campaign over the period of 2 weeks were local law enforcement agencies target distracted drivers. This can be done through the use of targeted billboards, visible law enforcement as well as a sting operation during peak hours.
The clustering map showing areas of danger based on the longitude and latitude can be much easier to understand versus the previous A-Priori analysis. This is because only a single location is being put into perspective at a time with association analysis. By viewing a graph as sen in figures 5 and 6, we can come to a greater understanding of potentially hazardous areas in New York. Spots highlighted in red in figure 6 have a high frequency of accidents and should be avoided when necessary. It is also up to local authorities and city ordinance to determine the best path in fixing areas that are prone to creating accidents.

6. Conclusions

In this paper, three methods were used to analyze the data set containing traffic records from New York. During the association analysis, it was discovered that lowering the support threshold to 0.05 was necessary due to the large number of possible combinations. This threshold allowed the top results between types of accidents and borough location to be shown. The K-means clustering method of analyzing the data has thus far been inconclusive. Attempting to use this method to determine if there was a correlation with certain longitudes or latitudes being particularly accident prone showed as inconclusive. However, another method using heat maps

References

Joan Lowy, Associated Press. (2014, May 29). Traffic accidents in the U.S. cost $871 billion a year, federal study finds. PBS NewsHour. https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds