Case Study on Car Crashes in New York

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

[1] Quantitative Imaging and Nanobiophysics Group, MRC Laboratory for Molecular Cell Biology and Department of Cell and Developmental Biology, University College London, Gower Street, London, WC1E 6BT, United Kingdom
[2] tesst

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 popular cities that is 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.

1. Introduction

1.1 Domain introduction

1.1.1 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.

1.1.2 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, associate rule mining, clustering, classification, regression.
1.2 World Health organization reports that approximately one million three hundred fifty thousand individuals lose their existences and die yearly because of road accidents.

2. Methods

To analyze the dataset, three methods were used. These methods consist of association analysis using the A-Priori algorithm, clustering with K-means clustering and … Depending on the algorithm used, the size of the dataset 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 dataset was used in applicable instances where the computing power did not exceed the limitations.
2.1 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 correlations and co-occurrences exist. Generally the A-Priori algorithm performs slower than other methods on large datasets. 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.
2.2 Clustering with K-Means

Conclusions

ACKNOWLEDGEMENTS

References