Car accident severity in Canada 1999–2017

Applied Data Science Capstone

Introduction

On every day thousands of people die from traffic accidents. Often, young people lose their fruitful lives at dramatic car crashes. The costs of fatalities and injuries due to accidents on society are high, therefore governments and the industry invest millions of dollars into the safety of cars, roads, and infrastructure. However, as the world population rises and rises, roads stay congested, and there are still too many unnecessary car accidents and death cases therefrom.

Nowadays, rich data is available, and modern machine learning tools can be applied to analyze these data sources. In order to make intelligent decisions, governments need to know where money is best invested. Thus, people started to determine main factors causing severe traffic accidents and revealed patterns and clusters. This includes weather conditions, road conditions, time, driver's age, and many additional factors. Furthermore, to reduce the number of severe car crashes, it is essential to have precise models that are capable to make accurate predictions. A specific forecast model helps decision-makers to decide between different safety invest as it anticipates by how much an action may decreases the number of accidents. It could also be used at extreme weather situations to alert drivers.

In recent years, having more computaational power and more elaborated machine learning methods at hand, a lot of research has been done in this field. For instance, Abdelwahab et al. used neural networks to predict severities [1], Chong et al. applied several different machine learning methods to the problem [4], or AlMamlook et al. compared different machine learning algorithms to predict accident severities [2].

In this report, we consider the car accident severity in canada [3]. We have two main goals: first, we anaylze the data set, understand the relationships, and illustrate how the number of car collision in general and fatal collisions in particular depend on factors such as time, weather condition, or roadway surface. Second, we apply different standard machine learning methods to our data and compare their accuracy in predicting the severity of car crashes. The main technical problem in our analysis was the size and the imbalancedness of the data source: only a 0.02-fraction of the records belong to severe collisions. We overcame this issue by sampling data sets with a large bias towards fatal cases.

References

[1] Hassan Abdelwahab and Mohamed Abdel-Aty. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record*, 1746:6–13, 01 2001.

[2] Rabia Emhamed AlMamlook, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, and Abdulbaset Ali Frefer. Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 272–276, 2019.

[3] Transport Canada. National collision database. https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a, 2017.

[4] Miao Chong, Ajith Abraham, and Marcin Paprzycki. Traffic accident data mining using machine learning paradigms. *Informatica*, 01 2005.