Prediction of the Car Accident Severity for the City of Seattle

Coursera Capstone Project

Silviya Ninova

September 16, 2020

Introduction

In its "Global status report on road safety 2018", the World Health Organization (WHO) states that road traffic injuries took the life of about 1.35 million people.[1] It is the main cause of death for young people in the age group 5-29 years and the 8th leading cause of death globally.[1] Such a tragic situation often comes at a high emotional cost. The victims' families are left to cope with the fact that their loved ones have been suddenly taken away from them. Survivors often have to suffer the devastating impact lifelong injuries have on their physical and mental health, and financial status. At an economical level, this is translated as up to 3% of the annual gross domestic product going for costs related to road accidents.[1] It is thus of vital importance on global and local scale to have a strategy for road-accidents prevention, so as to save the lives of the generations to come.

High-income countries are reported to have an overall decrease in the fatal road accidents, as compared to low- and middle-income countries.[1] Among the former is the United States of America, which however individually shows the opposite trend. The fatality rate per 100,000 population did drop by 4.4 between the years 2005-2014. As of 2014, however, the numbers have been slowly crawling back up again with a peak in 2016.[2] Identifying thus the major factors leading to an increased accident severity is of paramount importance in the road traffic injury prevention. Potential victims could be saved from a hospital stay or even long-term damage, which would also influence positively the economical costs associated with these.

In the present report, I am going to investigate road accidents on a local scale, namely for the town of Seattle. The aim of this project is to determine the variables on which the accident severity depends and whether these can be used to predict the risk of a collision with injured people. Such knowledge is of paramount importance to the local authorities, who in line with the strategy for road traffic injury prevention, can take the necessary precautions and alert the local population for high-level collision risks.

Data

The data set used for this project is readily available online on the Seattle GeoData[3]. It contains information on all collisions provided by the Seattle Police Department and recorded by Traffic Records Groups from 2014 until 2020. The original data contains 194673 records and 37 features. Not all features in the original dataset, however, provide relevant information for this project, so I started with a selection of pertinent columns.

The target variable in this model is the `severity` of a accident. In the present data set it can adopt two values: 1 for *property damage only collisions* and 2 for *injury collisions*.

Features proven to be relevant in road-vehicle collision models are the weather conditions, the road (type, condition, etc.) and the time/date of the accident.[4, 5] This informations is

already present in the used database under several column values. The `weather` feature reports 10 different meteorological conditions concerning the presence of clouds (*overcast*, *clear*, *partly cloudy*), rain (*raining*, *sleet/hail/freezing rain*), snow (*snowing*), strong winds (*severe crosswind*), fog (*fog/smog/smoke*) or other (*blowing sand/dirt*, *other*).

Information on the road is stored in a variety of features. The road conditions are stored in the `road-condition` and report the presence of standing water, ice, snow, mud or oil on the road at the time of the collision. In addition, it is distinguished between wet or dry road. Another factor regarding the road infrastructure is the degree of illumination it has (`illumination`). In this case a differentiation is made between daylight, dusk, dawn and dark with street lights on/off/missing. It is rather easy to oversee a pedestrian/cyclist or even cars on a night drive when the visibility is poor. Whether accidents at intersections, blocks or alleys are more probable, is stored in the features `address-type` and `junction-type` Finally, information on the exact address is stored in `latitude` and `longitude`. This is useful for a map representation of the data, so as to understand whether some neighbourhoods or roads tend to be the scene of collisions more often than others. The precise address in `LOCATION` does not bring any additional information, so it is dropped from further analysis.

Another important factor is the date of the accident (`day-of-year`) within a specific year (`year`). This is indicative of how the seasons, day-length, holidays affect the severity of a collision. Moreover, since accidents are prevail at certain hours of the day, we extracted this information in a separate attribute `hour-of-day`.

Finally, I look into the different collision types (collision-type) and to what extent the collision is a result of high speed (speeding), inattention (inattention), represented as a binary attribute. All records with "not enough information" for the `EXCEPTRSNCODE` feature (5638) were dropped from the data set, due to missing important information.

References

[1] Global status report on road safety 2018. Technical report, World Health Organization, 2018.

[2] National highway traffic safety administration. Technical report, United States Department of Transportation, 2018. [Online; accessed 15-September-2020].

[3] Seattle GeoData. Collisions. `https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions?geometry=-123.320%2C47.452%2C-121.342%2C47.776`, 2018. [Online; accessed 16-September-2020].

[4] Antoine Hébert, Timothée Guédon, Tristan Glatard, and Brigitte Jaumard. High-resolution road vehicle collision prediction for the city of montreal. In *2019 IEEE International Conference on Big Data*, pages 1804–1813, 2019.

[5] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, page 33–42, New York, NY, USA, 2019. Association for Computing Machinery.