Predicting the severity of an accident — IBM Applied Data Science Capstone

Paul Moreira

October 2020

Contents

1 Introduction

1.1 Background

According to the World Health Organization approximately 1.35 million people die each year as a result of road traffic crashes, and it costs most countries 3% of their gross domestic product. Traffic accidents cause economic losses to people and their families, and not only material losses because in the event of death or injuries, they affect several other aspects of the community. Starting from the premise that traffic accidents can be prevented, it is imperative to know the main factors that can cause an accident in order to develop strategies for the government or some organizations to act.
We will analyze a data set consisting on all type of collisions from 2004 to late 2020 in Seattle, provided by SPD and recorded by Traffic Records. By applying some machine learning techniques, we will predict the possible outcome of a traffic accident in terms of fatality.

1.2 Problem

In Seattle, as it is for most cities in the world, it is necessary to implement initiatives to reduce the rates and severity of traffic accidents, so this project aims to predict whether the outcome of an incident is fatal or not based on available data.

1.3 Interest

Drivers, pedestrians, cyclists could benefit from the results of the prediction, because they could act under certain conditions that are very likely to cause an accident.
Government, traffic departments, police who need to take steps to create safe roads, build safer infrastructure, improve post-accident care for victims and raise awareness.

2 Data acquisition and cleaning

2.1 Data sources

A comprehensive dataset consisting of road incident records that occurred between 2004 and October 16, 2020 in Seattle, obtained from the Seattle Open Data Portal 1. This dataset contains 221,524 rows and 40 columns, which describe the details of each incident, including location, severity, type of collision, fatality, date and time, weather, and road conditions, among others.

2.2 Data preparation

Once the data set was downloaded, the analysis of the attribute information of the dataset was performed determining that, in the original form, the dataset is not suitable for data analysis. There are several issues with the dataset and it will be handled as follows:

Relevant features: The 15 most relevant features were selected so the rest of the characteristics were discarded. The kept features are:

- SEVERITYCODE: A code that corresponds to the severity of the collision.
- SEVERITYDESC: Description of the severity of the collision.
- COLLISIONTYPE: Collision type.
- JUNCTIONTYPE: Category of junction at which collision took place.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N).
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.
- WEATHER: A description of the weather conditions during the time of the collision.
- ROADCOND: The condition of the road during the collision.
- LIGHTCOND: The light conditions during the collision.
- SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)
- SEGLANEKEY: A key for the lane segment in which the collision occurred.
- HITPARKEDCAR: Whether or not the collision involved hitting a parked car. (Y/N)
- ADDRTYPE: Collision address type.
- INCDATE: The date of the incident.
- INCDTTM: The date and time of the incident.

Incompleteness: There are missing values for most of the features even in the label that we want to predict, therefore, in order to build a model and determine how the features interact with the accident severity variable, we determined that it is necessary to remove the missing data entries of more than 2 key variables.
Missing values: We identified the missing values for each characteristic and analyzed them individually, for most of the characteristics it was necessary to replace the missing values with the highest frequency. In the case of INATTENTIONIND, SPEEDING, the dataset contains only the positive values, so for the rest it was replaced by the value 'N'.
Correct data format: We proceed to convert data types to the proper format, as it is the case with datetime features.
Normalized: For the UNDERINFL characteristic, it was necessary to normalize the values unifying 'N' with 0 and Y with 'S'. For the function WEATHER, LIGHTCOND, the values 'Other' were unified with the 'Unknown'.