Predicting the Severity of a Car Accident

By Iker Sedano Mattheus

October 6, 2020

1. Introduction

1.1 Background

Let’s say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to. Well, this is exactly the porpoise of this project.

1.2 The problem

Data of previous car accidents is needed, the accident severity should be labeled and accompanied with different attributes like time of the accident, weather, road condition, … etcetera. This project aims to predict the severity of a car collision given some conditions to warn drivers of potential heavy traffic in their ways.

1.3 Interest

All kind of drivers could benefit from the results of the project. Drivers that want to avoid any inconvenient in their trips could plan alternative ways to avoid long traffic waits. Also, could improve the safety awareness of the drivers motivating them to drive more carefully or even to cancel their trip if the conditions are bad or if the chances to encounter with a severe accident are high. In addition, this predicting model could be integrated in existing applications for drivers like Waze, Google Maps or Apple Maps, these applications can take information of weather and road conditions and make predictions and give the information to the driver at every stage of the trip.

2. Data acquisition and cleaning

2.1 Data source

The data used to develop the predicting model is gathered from the Seattle Police Department (SPD) and can be found here. The data set represent a record of accidents, the shape of the data frame is 194,673 rows and 38 columns. 37 of the columns are the attributes or the independent variables and 1, the severity of the accident, is the labeled data or the dependent variable.

2.2 Feature selection

The dependent variable or the feature that is going to be predicted is the severity of the accident (SEVERITYCODE). It uses the following numerical code to categorize the severity of the car accidents:

- 0 – unknown
- 1 – property damage (no injuries)
- 2 – injuries
- 2b – serious injuries
- 3 – fatality

Giving a quick look to the original data set can be seen how all accidents are labeled in two of the categories described above: property damage (1) and injuries (2). There are 136,485 samples for the first one and 58,188 for the latter.

Now that the first look to the dependent variable is made let's see the independent variables that are going to be used for the developing of the model. As it is mentioned before, the goal of this project is to predict the severity of an accident given certain conditions. So, exploring the 37 given features in the data set the conditions that can fit the model are: weather, road condition and light condition. Let's do a preliminary analysis of this features. For WEATHER the labels are clear, raining, overcast, unknown, snowing, other, fog/smog/smoke, sleet/hail/freezing rain, blowing sand/dirt, severe crosswind, partly cloudy. Next, for ROADCOND the labels are dry, wet, ice, snow/slush, standing water, sand/mud/dirt, oil, unknown, other. Finally, for LIGHTCOND the labels are daylight, dark – streetlight on, dark – streetlight off, dark – no streetlights, dark – unknown lightning, dusk, dawn, unknown, other.

These features are selected because are the ones that can be implemented gathering data from the real world. So, when some for each feature one of this labels should be the input and the model should predict the severity of the accident the driver may encounter, giving the user the ability to make a proper decision about its trip.