Predicting the Severity of getting in a Car Accident through a Machine Learning Model

Emilio Padrón Molina
emiliopm1997@gmail.com
IBM Data Science Professional Certificate

September 23, 2020

Introduction

Accidents happen around the world all the time, causing unnecessary deaths and injuries. Some of them can not be avoided such as natural disasters and lightning strikes; however, there are other types of accidents that can be prevented by technology just like smoke detectors foresee potential fires. Car accidents have been one of the most common types of accidents through out the XX and XXI centuries, and even though there are global and local efforts to decrease this statistic, the numbers are still alarming.

Giving that this problem concerns a big part of the population, some people might have asked themselves: Is there a way to predict a car accident before it happens? If there is, companies that use navigation systems, such as Uber, Google, and Rappi, could take advantage of this and implement a model that could not only optimize traveling time, but also prevent a potential accident.

In this report, a classification model will be built to try to predict the risk of getting in a car accident. Nevertheless, before building a model, it is important to consider the reasons why car accidents happen such as traffic, weather, or poor road conditions, and how this data can be obtained.

Data

Relevant data to carry out this project was obtained from the Seattle Police Department who recorded car collisions in their city. These records go from January 1st, 2004 to May 20th, 2020 and are composed of 194,673 incidents. In the best case scenarios, incidents contain 38 characteristics including the severity of the accident, weather conditions, and type of collision. Nevertheless, it is worth mentioning that neither all the features are significant nor all the incidents contain a full description (there are many features are missing).

After analyzing the dataset carefully, one can say that the most significant characteristic that relates to predicting the possibility of a car accident is the "Severity" column which only contains the values: 1 and 2. Number 1 implies that the accident only resulted in property damage while number 2 indicates that there was an injury involved and, therefore was more severe. Given this, a classification model can be built to predict the "Severity" through the most significant variables in the dataset.

At a first glance, it is easy to notice that many of the incident's characteristics are consequences of the accident itself such as "number of people involved", "number of pedestrians in the accident", and "number of vehicles involved", and therefore, are directly related to the "Severity". Given the redundancy of this information, these sort of columns can be dropped from the dataset. Additionally, there are other variables that work as keys of incidents but have evidently no relation to the severity of the accident, thus these could be deleted as well. After this filtering, the dataset that will be used for an exploratory analysis contains 17 features plus the target.