Coursera Capstone Project: Applied Data Science

1 Introduction

Car crashes and road accidents could be considered an old topic. Yet, with the progress of the technology involved and the capabilities of these sophisticated machines, it is ever more important to have tools and means available to mitigate their occurrences, as well as their implications and consequences for the people involved. Thanks to the advancement of technological and analytical tools in the last two decades we are now able to better understand how crashes happen. This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly. Many attempts have been taken by many professionals, scholars and government agencies to provide produce these models; each with different goals, ways of measuring success and precision of their results.

2 Business Problem

Predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on the data available and how well the system or model has been defined. However, if the dataset's features are clearly defined and if there's a thorough description of how this data is collected –as is our case– we have much better chances of arriving at a usable model. In this particular dataset, the severity for a crash is classified under four different categories: non-injury, minor, serious and fatal. Thus, the main objective of the project will be to create the machine learning model should be able to predict accident "severity. This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

3 Data

The dataset that we will use comes directly from the Organization SDOT Traffic Management Division, Traffic Records Group and is made available through the dataset (its CSV version).
It contains data from all types of collision since of the year 2004 until the present day and it's automatically updated. The dataset has an initial total of 655,697 samples. Each with 89 different features. The feature we want to predict is Severity and it is a numerical feature that takes the values 1 and 2 to indicate that the crash was of class Property Damage Only Collision, or Injury Collision, respectively. The remaining features are both numerical and categorical. Additionally, there's also two geo spatial features (X and Y) used to indicate longitude and latitude of the crash site. The file metadata is a printed PDF expanded catalog created for this project specifically, also It has a complete list of all features in the dataset and their meaning.