Car Accident Severity Report

Applied Data Capstone - Coursera

Diego Talavera

October 11, 2020

Table of Contents

1 Introduction

1.1 Problem Description

Driving conditions can vary widely based on a very large number of factors: from the amount of traffic, to the weather conditions, lighting of the road and even the time of day. Thankfully, nowadays we have a very vast amount of information available from several governmental agencies that can offer some insight into the effects of the conditions mentioned earlier on chances and severity of an accident.
Developing a tool to predict the likelihood and severity of accidents can help to plan shorter and safer trips, from a daily commute to a month-long road trip. This tool could be of interest for government agencies and insurance companies in order to reduce the time taken to classify the accidents and in the case of the latter, it can make it easier and faster to determine the payout required for the accident.

1.2 Description and use of data

In this section, the data will be described, as well as some assumptions needed to simplify the analysis. The methods to be used are also presented here.

1.2.1 Data descriprion

- **Collisions information:** A dataframe obtained with various attributes that add up to a severity code in order to quickly assess the accident. This severity code is the variable that this work aims to predict taking into account the *relevant* attributes of the dataframe.
- The dataframe originally consists of 38 Columns and 194 673 rows
- **Data Cleanup:** This dataframe needs to be cleaned and maybe transformed depending on the selected machine learning tools. For example, some methods cannot handle 'str' variables andccan only work with integers or floats. Given that some of the information is in the dataframe are text strings, will have to be translated to a numeric value to make the data handling easier.

1.2.2 Data usage, assumptions and observations

In order to make the data analysis less complicated, some assumptions and considerations are made after a brief look at the data and metadata provided:

- The attributes presented in the metadata does not perfectly match the ones in the dataframe. This reduces the complexity of the analysis but it also reduces the accuracy and relevance of it.
- One of the examples of the previous point is the "severity code"(SEVERITYCODE) of the dataframe: While in the metadata there are 5 different possible codes, in the dataframe only two of these codes are used. This reduces the problem to a binary classification analysis and thus, none of the other labels presented in the metadata will be considered.
- The data will be split into training and test data to develop and tune the model.
- Data not described in the metadata will not be taken into account, therefore removed from the dataframe

1.2.3 Methods

First, Pandas will be used to create our starting dataframe from the acquired CSV file with the traffic accidents information. A few preliminary statistical analysis will be performed and then the data cleaning will also be performed using pandas, numpy, scikit learn and some other python libraries where applicable.
Depending on the results of the preliminary data analysis, a machine learning algorithm will be selected, trained and tuned, in order to develop a model that can predict the severity of the accidents. This method will be a supervised learning algorithm and likely a binary classification algorithm.