Coursera Capstone Project

IBM Applied Data Science Capstone

Car accident severity

By: Dr Boaz SAINT-LOUIS

Sept 2020

Table of Contents

1. Business Problem
2. Data
3. Methodology
4. Results
5. Discussion
6. Recommendations

1. Business Problem

In order to complete the IBM Applied Data Science Certificate, we want to explore the Collisions dataset, from Seattle SPOT Traffic Management Division and provided by Coursera, to evaluate which characteristics or features can be used to predict accident "severity".

We will use machine learning and data science technics to develop a model to predict car accident "severity" in terms of human fatality. **The target audience** is **road users in Seattle** and, in particular, **those responsible for the implementation of traffic policies in transportation division**. Our goal is to help this audience to:

(i) Have a better transparency over the key accident “severity” drivers

(ii) Incorporate these factors in their travel decision making process

(iii) Act on them in order to reduce the risk of being involved in a severe accident in terms of human fatality

2. Data

The Collisions **dataset consists of data on all types of collisions** (e.g. Bicycle, Car, Collisions, Pedestrian) in Seattle. All collisions data are provided by SPD and recorded by Traffic Records from 2004 to Present. The data are updated on weekly basis. The current version contains 194,673 observations and 38 attributes or features. Given that our goal is to determine the key factors that cause collisions and the level of severity, **we will extract**, for example, the following features: weather conditions, road conditions, speeding, light conditions and other relevant factors as **predictor variables**. As for accident severity, in terms of human fatality, we will use **the severity (SEVERITYCODE) as dependent variable or target**. For example, a code of 3 is assigned to accident classified as “fatal” while a code of 2 is assigned to accident with “injury”.

The dataset contains also several missing values. We will identify these missing values and explore different methods to handle (remove or replace them) them. In addition, given the dataset size is quite substantial, they are several features (e.g. XCEPTRSNCODE, SDOT_COLCODE) that are irrelevant for our analysis. We will identify these features by using different statistical technics and exclude them.