CAR ACCIDENT SEVERITY

Capstone project

Week 1

2. Data

2a. Origin of data

Based on the process (CRISP-DM), we can download data that will be utilized by SDOT Traffic Management Division, Traffic Records Group, Seattle, United States).

In the file (Data-Collisions.csv) the column named "Severity Code" consists of two values:

- 1 = property damage
- 2 = injury

Also, other columns report various accident conditions such as: location, weather, light, road, types of collision, etc.

2b. Data balancing

We observe that the targeted variable "Severity Code" has more references to value 1 = property damage than to value 2 = injury.

So we have to balance the two values so as not to be led to wrong conclusions.

Sampling is a widely used technique to address this issue. It consists of removing random observations from the majority class to prevent the dominant signal of the learning algorithm (under sampling) or the accidental repetition of observations from the minority class to amplify the signal (over sampling). 2 Under sampling, a large amount of data will be lost, which can later be used to predict severity. Therefore, the hyper-sampling technique is preferred and applied in this project.

2c. Clearing the data

We notice that there are many empty "NaN" cells, which we will replace with the values that appear in the maximum range.

In addition, in order to develop regression models, certain variables must be converted to index variables.

Some other columns contain data values regardless of the analysis we want to perform or are overshadowed by other data, while some will be scrolled and renamed.

2d. Selected Independent / Variable Forecasts:

1. LONGITUDE
2. LATITUDE
3. INDIVIDUAL (total number of people involved in the conflict)
4. VEHCOUNT (total number of vehicles involved in on collision time)
5. JUNCTIONTYPE (category of the intersection at which the collision happened)
6. WARNING (whether or not the collision is due to carelessness)
7. WEATHER (weather conditions on collision time)
8. ROADCOND (condition of the road on collision time)
9. LIGHTTCOND (lighting conditions on collision time)
10. SPEED (whether speed was a collision factor or not)