IMPROVING ROAD SAFETY IN FRANCE

Andrea Mazzoleni

SUMMARY

INTRODUCTION

1. Background

Road safety has always been one of the biggest problems in the world. Despite the technological progress, road accidents are still one of the main causes of death for people today: in fact, according to the CDC[1], *"Road traffic crashes are the world's leading cause of death for children and young adults 5-29 years of age"* and *"Each year, 1.35 million people are killed on roadways around the world"*. Although cars and other road vehicles have achieved high levels of reliability and safeness, they cannot (yet) completely replace humans in the management of driving dynamics. So, any kind of human distraction is enough to defeat all sorts of safety controls and cause an accident. Moreover, the number of people driving vehicles has grown considerably compared to previous years and, as we know, the greater the number of variables involved, the greater the extent of the problem.

That said, road safety is the main form of road accident prevention and is therefore always in the sights of local administrative authorities. The maintenance of the road surface, the provision of adequate acoustic and visual signs, the patrolling of high-risk areas are all measures aimed at reducing the number of accidents recorded each year. Unfortunately, these measures are often very expensive and are not always implemented optimally.

2. The business problem

As part of this project, I play the role of a French administrative authority that is trying to increase the level of road safety in its cities. The goal is to identify strategic points for strengthening road accident prevention measures (e.g. improve road sign, place road bollards and speed control systems, and so on). I have to understand which combination of environmental and meteorological factors determine the severity of an accident, so that

[1] Centers for Disease Control and Prevention

the authorities can intervene in a specific and targeted way on highways, intersections, crossroads and any other kind of road infrastructures.

The focus is mainly on the points at greatest risk, i.e. those in which a hypothetical accident is most likely to have serious consequences for people (disabling injuries or death).

This is where Data Science comes in. Through the analysis of the data collected over the last few years, I want to build a model capable of predicting the severity of an accident based on a limited set of attributes. The attributes of interest will be those of a meteorological and environmental type: morphology of the road, atmospheric conditions, lighting conditions, traffic regime, etc.

DATA

3. Data description

The data processed in this project comes from https://www.kaggle.com/ and are part of the *"Accidents in France from 2005 to 2016"* dataset. The dataset is divided into five different subsets:

- *Caracteristics* - contains information about the place, date, weather conditions, type of intersection and type of collision related to the accident.
- *Places* - contains all the information related to the place where the accident took place.
- *Users* - contains information about people involved in the accident.
- *Vehicles* - contains information about the vehicles involved in the accident.
- *Holidays* - contains information regarding public holidays year by year.

By considering each of its subsets, there is a total of 54 attributes inside the dataset. Obviously not all of them are needed to perform the analysis. In fact, as mentioned in the previous paragraph, the analysis is focused on environmental and meteorological factors to predict the severity of an accident.

That said, the first step in approaching the data has been to exclude non-relevant subsets and select the attributes of interest from the remaining ones.

4. Feature selection

Each subset has been explored separately from the others to better highlight the attributes and to process each of the related *csv* files with different methods. In particular, the *Caracteristics* subset was poorly structured and it was necessary to adopt unconventional extrapolation methods. After the preliminary analysis, two of the five subsets were excluded from the project: *Vehicles* and *Holidays*.

From the remaining ones I have selected only a few attributes, which are described below.

From the *Caracteristics* subset I took 4 attributes:

1. *Num_Acc* - unique identifier for the accident. I am going to use this attribute to join the subsets into a single dataframe.
2. *lum* - lighting conditions.
3. *int* - type of road intersection.
4. *atm* - atmospheric conditions.

From the *Places* subset I took 4 attributes:

1. *Num_Acc* - unique identifier for the accident. I am going to use this attribute to join the subsets into a single dataframe.
2. *catr* - category of the road.
3. *circ* - traffic regime of the road.
4. *surf* – surface condition of the road.

From the *Users* subset I took 2 attributes, which are:

1. *Num_Acc* - unique identifier for the accident. I am going to use this attribute to join the subsets into a single dataframe.
2. *grav* - the physical consequences reported by the people involved in the accident.

5. Data preparation

Once the features of interest were selected, I had to merge them into a single dataframe. However, while each row in the first two dataset was uniquely identified by the *Num_Acc* attribute, in the third dataset there were more entries with the same value of *Num_Acc* (one for each person involved in the accident). So, to successfully merge the subsets, I had to make *Num_Acc* a key in the third one. This was done simply by maintaining the row with the highest value of the *grav* attribute for each accident and removing the others.

This is also the way I built the severity attribute: in fact, the severity of an accident coincides with the most serious case among the injuries reported by the people involved. So, after this operation the *grav* attribute has become our severity attribute.

Finally, the three subsets were merged into a single dataframe with the following attributes: *Num*_Acc, *grav, lum, int, atm, catr, circ, surf* (for a total of 8 attributes).

I am not going to describe each attribute in detail, but it is important to indicate what values the *grav* attribute can take:

- *grav = 2*: Death.
- *grav = 3*: Serious injury.
- *grav = 4*: Light injury.

6. Data cleaning

There were 839985 rows inside the created dataframe at this point. Unfortunately, many entries had *NaN* values and some actions had to be taken. Since all the attributes are categorical, I chose to not replace the missing values but simply remove them (i.e. the rows in which they were present).

After this operation, the rows inside the dataframe were 524263.

Finally, it was necessary to convert each column type to a numeric one, to facilitate the operations that will be performed on the data.

	Num_Acc	grav	lum	int	atm	catr	circ	surf
0	200500000001	4	3	1	1	3.0	2.0	1.0
1	200500000002	3	1	1	1	2.0	0.0	1.0
2	200500000003	3	3	1	2	2.0	0.0	2.0
3	200500000004	4	1	1	1	3.0	2.0	1.0
4	200500000005	4	3	1	3	3.0	2.0	2.0

Figure 1 - The final dataframe