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

METHODOLOGY

7. Exploratory data analysis

As already mentioned, the goal is to find a relationship between environmental and meteorological factors and the severity of an accident. At this point, it is not clear if any of these factors (or a combination of them) are related to the severity attribute or not. So, to get more insights about the data, an in-depth analysis was carried out.

At first, the correlation matrix for the previously created dataframe was built. Unfortunately, none of the features showed a strong correlation with the severity attribute.

Then, the distribution of the values of each attribute in relation to the severity level was analysed. In this way, I could understand the extent to which each feature varies based on the severity attribute. This was useful to confirm (or reject) the assumptions made about the relationships between the data before moving on to model development.

The considerations drawn from the analysis performed for each attribute are listed below.

- ***Type of intersection ("int" attribute)*** - The histogram shows that the distribution is dominated by a single value, that is "No intersection". This means that most of the recorded accidents occurred outside intersections. However, this value is more dominant as the severity value decreases (i.e. when the severity of the accident increases, since a value equal to 2 corresponds to the maximum severity and a value equal to 4 to the minimum).

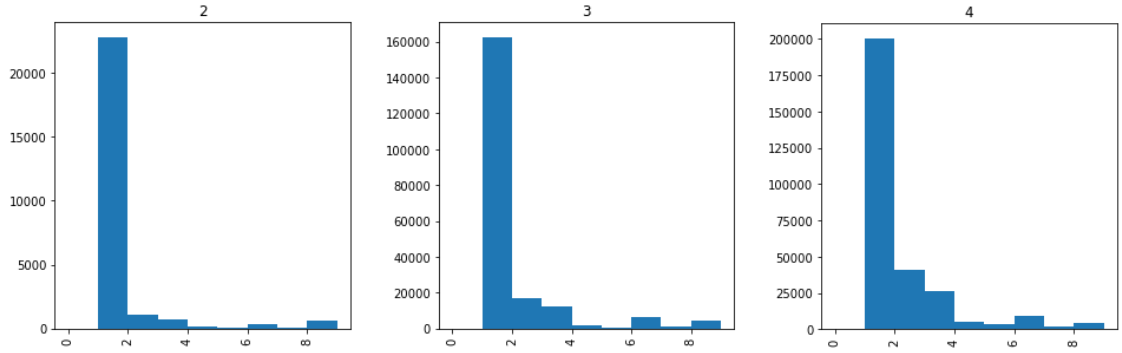

Figure 2a – Type of intersection distribution

- ***Atmospheric conditions ("atm" attribute)*** – Also for this attribute there is a dominant value, which is "Normal". The distribution of the other values varies moderately, showing a slight tendency to detect more serious accidents when atmospheric conditions are not optimal.

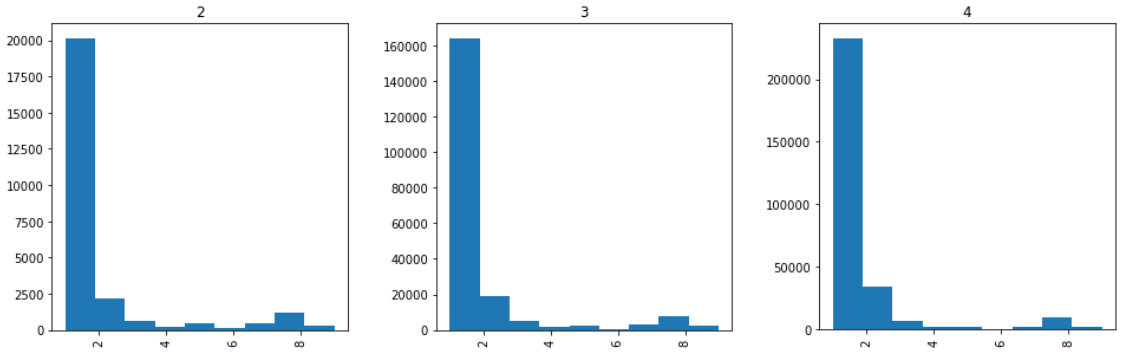

Figure 2b – Atmospheric conditions distribution

- ***Category of road ("catr" attribute)*** – In this case the distribution varies considerably from low to high levels of severity. The histograms show that accidents with light injuries are distributed uniformly along the main road categories, while more serious accidents happen mostly on Departmental roads.

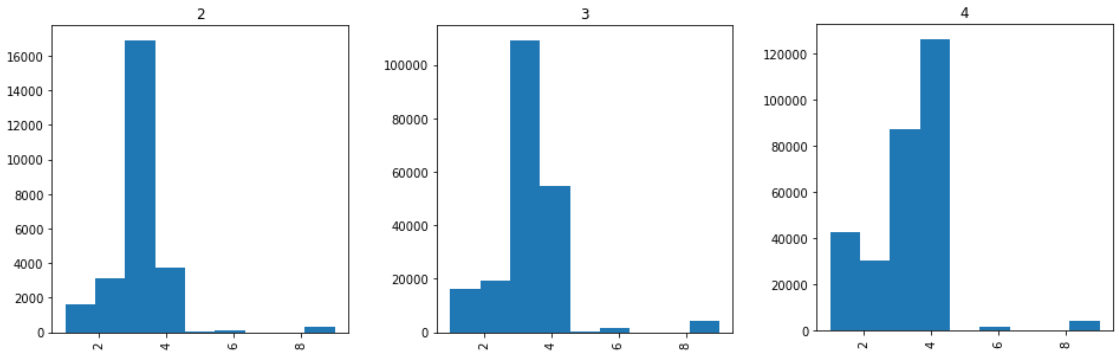

Figure 2c – Category of road distribution

- ***Lighting conditions ("lum" attribute)*** – The first thing to note is that there is a dominant value, which is "Full day". Furthermore, we can draw an obvious conclusion from the graphs: the accidents become more serious while lighting conditions worsen (see the distribution of the value equal to *3*, i.e. "Night without public lighting).

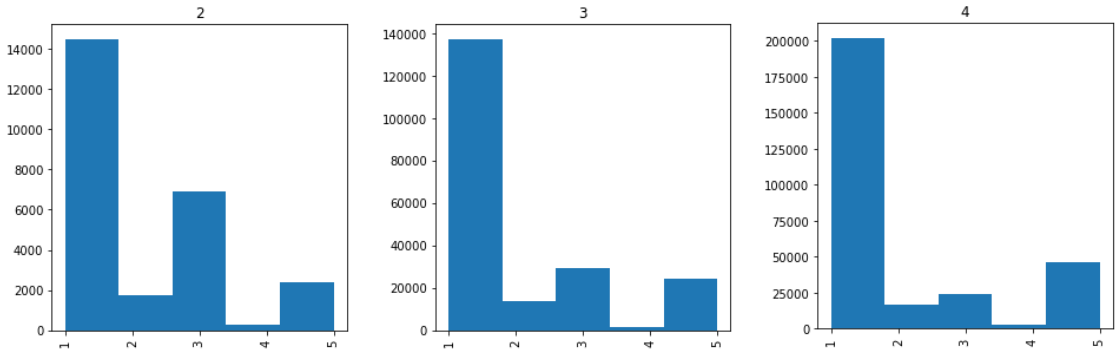

Figure 2d – Lighting conditions distribution

- ***Surface condition ("surf" attribute)*** – The distribution of this attribute is not significantly varying at different levels of severity. Therefore, the surface conditions do not seem to be related to the severity of the accident.

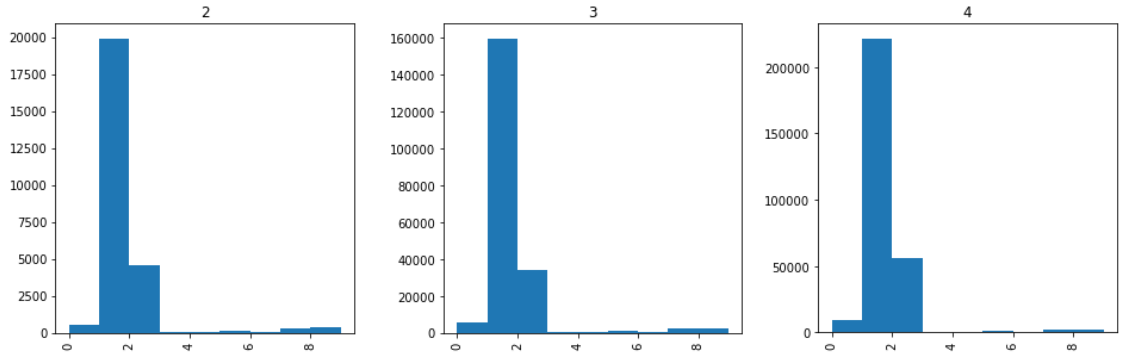

Figure 2e – Surface condition distribution

- ***Traffic regime ("circ" attribute)*** – Similarly to the first case analyzed, there is a dominant value which becomes more dominant as severity increases. This means that serious accidents happen mostly on roads with separated carriageways, while less serious accidents are more uniformly distributed along different types of road.

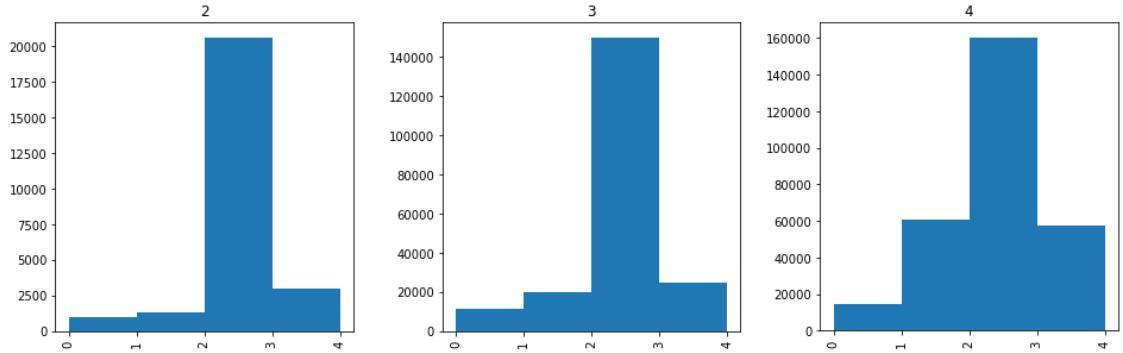

Figure 2f – Traffic regime distribution

The analysis showed the only factor that has no significant relationship with the severity of an accident is the *surf* ("Surface condition") attribute. For this reason, it has been removed from the final dataframe and the same was done for the *Num_Acc* attribute, as it is no longer needed at this point. All the other attributes have been considered suitable to be used in the model building phase.

8. Model building

The analysis phase showed that the assumptions about the relationship between the selected attributes and the severity of the accident were well founded. The next step was to choose an appropriate model for the data. Clearly, not all types of model and not all machine learning techniques are suitable for modelling the data in our possession.

After the last update, we have a total of 6 attributes inside the dataframe:

- 5 feature attributes (*lum*, *int*, *atm*, *catr*, *circ*), which are categorical variables with discrete values.
- 1 target attribute (*grav*), which is a categorical variable with discrete values.

That said, a supervised learning approach is needed to build a suitable model for the data.

In this context, I decided to build four different models using different classification machine learning techniques. Before proceeding with the creation of the models, however, it was necessary to perform some additional operations on the data.

- The feature data was standardized, giving them zero mean and unit variance.
- To properly train and test the predictive capability of each model, the data was split into two subsets:
    1. ***Train set*** - about 80% of the data.
    2. ***Test set*** - about 20% of the data.

The developed models are described below.

- **K Nearest Neighbors (KNN)** – The first model was created using one of the most known classification techniques, which is based on the distance between data points and whose purpose is to classify by identifying the *K* nearest points. The most important step while building this model was the choice of the value of *K*. Obviously, the goal is to maximize the model accuracy, so multiple models with different values of *K* were built, then their accuracy was compared. The best model (the one with the highest accuracy) was the one with a *K* value of *16*.

- **Decision Tree** – The second model is based on a well-known procedural algorithm which requires various hierarchical steps in the classification process. Each step involves the addition (*induction*) or the removal (*pruning*) of one or more decision branches and nodes from the model. The outcome is a tree, through which predictions on new data can be obtained.

- **Logistic Regression** – The third model was built by running an algorithm often associated with binary independent variables. Despite this, it can also be applied to multiclass problems such as the one subject of this paper. Overall, Logistic

Regression can be thought as an evolution of the Linear Regression where the independent variable is not continuous.

- **Random Forest** – The last model can be seen as an upgrade of the second one. In fact, Random Forest algorithm builds a large number of decision trees and make them work together as an ensemble. The *maximum depth* for this model was chose by comparing different models' accuracy, as was done for the first model. So, the best model was the one with a *maximum depth* value of *12*.

In addition to those listed, an attempt was made to develop a fifth model, based on the **Support Vector Machine (SVM)** algorithm. Unfortunately, since this type of model is kernel based and the fit time scales at least quadratically with the number of samples, it was really impractical and was removed from the project.

9. Model evaluation and results

To evaluate the predictive capability of each model the ***Test set*** was used. Then, a set of *scoring parameters* was formulated using the predicted variable $\hat{y}$:

1. ***Jaccard Index***
2. ***F1 score***
3. ***Recall***
4. ***Precision***

The results were collected in a new dataframe.

	Jaccard index	F1 score	Recall	Precision
KNN	0.421077	0.580856	0.593355	0.580114
Decision Tree	0.445565	0.600144	0.621642	0.588577
Logistic Regression	0.314139	0.403388	0.559812	0.468331
Random Forest	0.446262	0.600677	0.622624	0.595646

Figure 3 – Scoring parameters

As you can see from the picture, the Random Forest model has the highest values for all the calculated parameters. However, the scores for this model are really close to those of the Decision Tree case: this is not surprising, since the two algorithms are closely related to each other. The KNN model also achieved good results but slightly lower than those of the previous ones. Finally, the Logistic Regression model shows the worst results as

expected: although it is applicable in the present context, it performs better in contexts characterized by binary dependent variables.

Therefore, the Random Forest model represents the best model to fit our data.

CONCLUSION

10. Discussion and conclusions

The goal of this study was to find a relationship between the severity of an accident and a set of environmental and meteorological factors, so that strategic points alongside French roads could be selected to establish new road accidents prevention measures. Although some of the factors considered showed a strong relationship with the severity attribute, for others a poor incidence was found in this sense. This fact forced me to consider only a few of the available features while developing the predictive model, so the results obtained by calculating the *scoring parameters* were not very high.

Despite this, the project showed that the hypotheses were well founded, as the relationship exists and the model can predict the severity of an accident with a good level of accuracy. This is a great achievement, as there are generally other factors that are considered more incisive in determining the severity of an accident: think, for example, to the absence of a seat belt or the drunkenness of the driver of the vehicle. Being able to determine the severity of an accident through factors strictly linked to the territory allows to adopt very specific and selective prevention measures.

Furthermore, the model lends itself to be reworked by introducing new features into the data. Factors such as the slope of the road or the width of the carriageway were not present in the dataset but could make a significant contribution to the overall accuracy of the model.

In conclusion, road safety is one of the most important issues in the modern world and therefore requires the use of tools and knowledge aimed at continuously improving it.

This work is an example of how data analysis can help administrative authorities to make decisions and implement preventive measures in a targeted manner.