Predicting Severe Car Accidents

Nicolas Auvillain

October 24, 2020

1 Introduction

1.1 Background

Traffic collisions have an enormous cost. Most importantly, thousands get killed on the road every year. Beyond that, there are severe injuries, hospital bills, insurance mechanisms, lawsuits. Traffic data collected by law enforcement officers is publicly available. It is beneficial to all to identify patterns that can lead to serious injuries or death.

1.2 Problem

Which collisions are likely to lead to death or severe injuries? Those are the ones we need to avoid as a priority. Which factors are relevant in assessing that risk?

1.3 Interest

Everyone is interested in reducing the risk of death or severe injuries on the road. But more particularly, government officials, car makers, and even app developers. One could imagine Google or Apple warning drivers that certain conditions make severe accidents more likely.

2 Data Acquisition and Cleaning

2.1 Data Sources

The French government publicly provides data in the form of separate CSV files. We are focusing on data for the year 2018. The files are “Attributes”- the conditions in which the collision occurred, like the amount of light, weather conditions, and the state of the road; “Users” which contains data about the person(s) involved in the collision, like severity, age, gender; and “Locations” which describes the place where the collision occurred.

The data is to be found here: https://www.data.gouv.fr/en/datasets/base-de-donnees-accidents-corporels-de-la-circulation/

2.2 Data Cleaning

The 3 datasets were merged into one table. This was easy as each record contains a unique Num_Acc key that describes a specific collision.

One collision may involve multiple people, for instance a driver, a pedestrian and a passenger. I have decided to keep them all as they may all play a role in the collision.

Looking at age ranges, I found no outliers. I removed the records with missing values, then standardized each column.

2.3 Feature Selection

I decided to keep the features that were more likely to matter in the analysis; for instance, personal information had to be scrapped.

Other features, like the amount of light, the state of the road (wet, dry, oily), the light (daylight, darkness with no light) as well as the type of road (one way, two way with no divider, etc) seemed to be most likely to have an impact.

I did not take the time of day into account as we have no data over the total traffic. Maybe there are more accidents at 6PM than at 3PM, but there are more vehicles on the road too; so accidents may not be more likely then.

3 Exploratory Data Analysis

3.1 Calculation of Target Variable

As we have no data over the total traffic at the time of each collision, we cannot predict a "likelihood of collision". We can only try and ascertain whether certain conditions are more propitious to severe collisions than others. So we had to recategorize values 2 and 3 (corresponding to death and severe injury) into 1 and values 1 and 4 (no or light injury) to 0.

The counts show some imbalance – there are many more records for 1 & 4 than for 2 and 3: so we will need to balance the dataset.

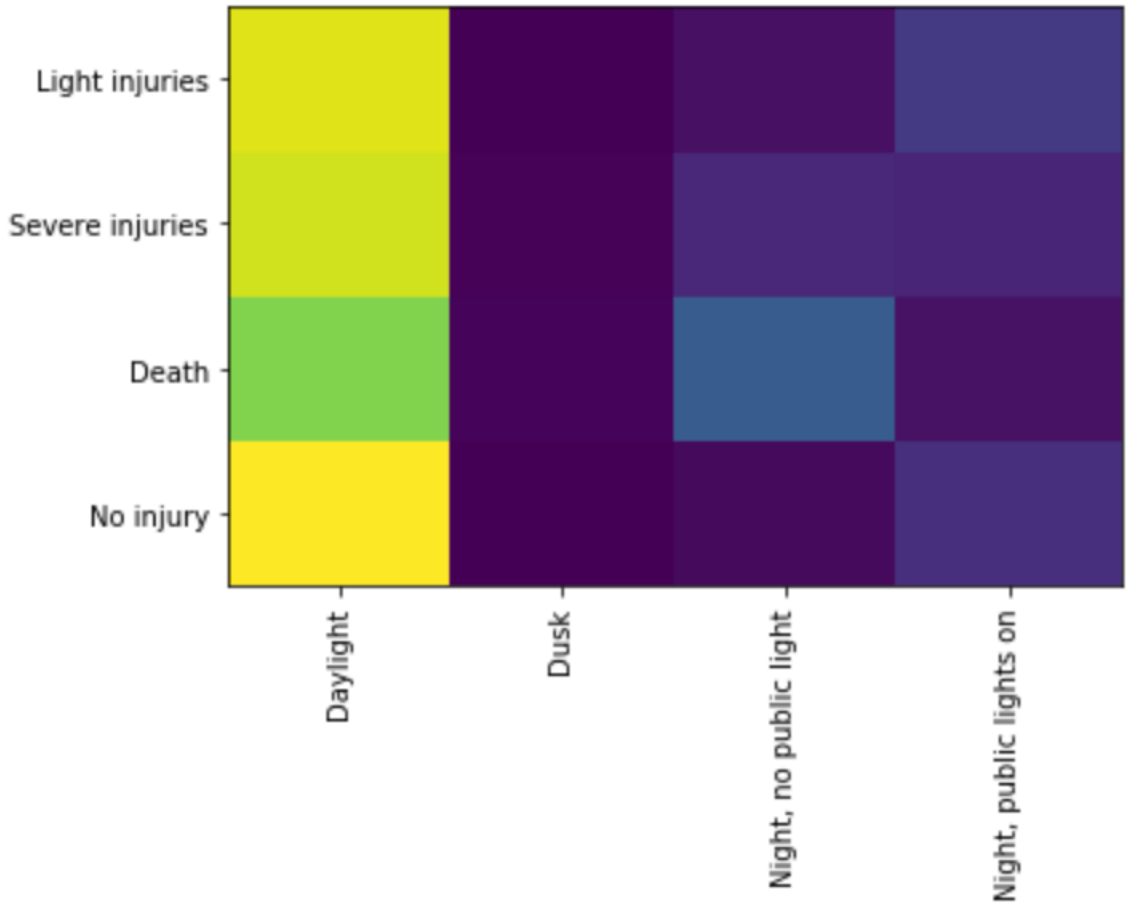
```
1    54248
4    50360
3    22169
2     3392
```

3.2 Relationships

3.2.1 Light & Injury severity

Intuitively, we feel that light conditions should have an impact on the amount of collisions.

Let's look at the heatmap:

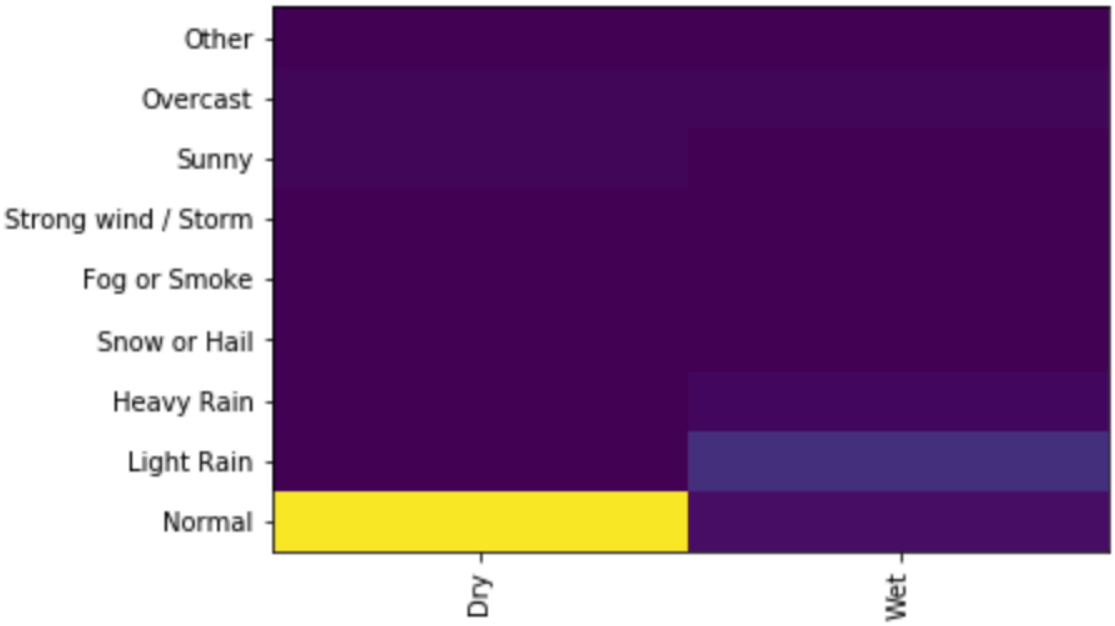


We see that most collisions happen in daylight, which may be due to the fact that most traffic occurs during the day. But we see that night collisions are more severe when there are no lights, which suggests that daylight sometimes provides the split-second reaction that is needed to avoid severe injuries.

3.2.2 Surface conditions and Weather

Let us look at how the state of the road and the weather affect the amount of collisions.

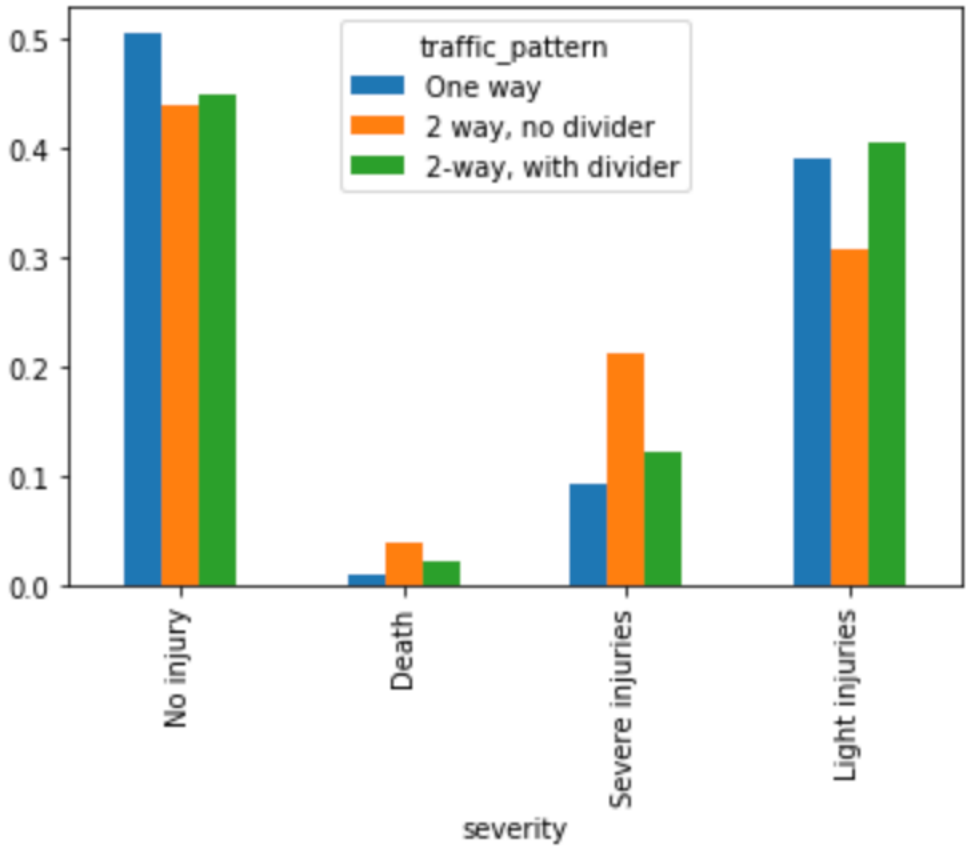
First we have to keep 1 record per collision (some collisions involve 2 vehicles, others 1). Then we generate the heatmap:



France is a temperate country, yet most collisions occur in perfect conditions: when the road is dry, and the weather is normal. The second highest is when there is light rain – probably because it can make the road slippery; but overall this suggest that people are careful in inclement weather, and most collisions are due to inattention during perfect conditions.

3.2.3 Traffic pattern vs Severity

Here is the table for traffic pattern vs. Severity:



We see that severe collisions occur prevalently on 2-way roads with no divider; as inattention seems the main reason for collisions, we can imagine that drifting onto the oncoming traffic lane is what causes them.

4 Predictive Modeling

I needed to balance the dataset first, as there are many more 'non-severe collisions' reported than severe ones. Then we standardized the dataset.

In this study, I will use supervised algorithms – to score the data depending on whether it is a severe collision or not. I've used Logistic Regression and Decision Tree clustering algorithms.

I partitioned the dataset in order to be able to train the algorithms, then test them to evaluate their performance.

4.1 Performance

Model	Jaccard_similarity_score
Logistic Regression	0.6777070063694267
Decision Trees	0.6837250164726554

Tweaking the parameters of the algorithms did not yield significant differences in the performance.

5 Conclusions

Performance in the model isn't very high, suggesting that the factors used are not good predictors for collisions.

What did seem to be a leading cause of collision was inattention – as most collisions occur in normal weather on dry roads. Also, night collisions are often severe if the lighting is inexistent. So, this suggests that improving lighting may be a way to save lives or reduce injuries.

6 Future directions

More data is needed to be able to draw conclusions; for instance, having traffic data would allow us to estimate the likelihood of a collision at a given time and place. Car data and sensors could provide information on safety distance compliance. Smartphones could allow for traffic data at the time of the accident, too. They could also provide information as to how distracted the driver is (for instance, if there is a phone call or a passenger is talking).
An app could then score the model in real time (say, at every change in road, light or traffic conditions, and when the driver gets distracted) and sound an alert.