

École polytechnique de Louvain

Modelling crash severity and safe-driving through vehicle dynamics

Author: **Marc-Henry COMANS**

Supervisors: **John LEE, Fintan MCGEE**

Readers: **Michel VERLEYSEN, German CASTIGNANI, Benoit OT-JACQUES**

Academic year 2019–2020

Master [120] in Computer Science

Abstract

Since the introduction of the first motored vehicles to the grand public, road safety has been an important concern for many actors. With the constant increase of new vehicle registrations around the globe and the growing interest in self-driving cars, understanding and preventing traffic accidents is becoming more and more crucial.

The objective of this thesis is to apply **Machine Learning** techniques to model car accidents to help us discover the main factors affecting their severity and creating new metrics measuring the danger of a road.

We will base our research on the history of accidents that occurred in Luxembourg between 2013 and 2017 and model them using **Decision Trees** and **Support Vector Machines**. We will see that the first can give us a measure of feature importance to consider when studying road characteristics in the context of traffic accident. The second will allow us to estimate a new measure of *danger* using the calibration method **Platt's scaling** in order to represent the risk of having *slight*, *serious* or *fatal* accidents when driving on a certain road.

Additionally, we will introduce the concepts of **Data Visualization** and practical applications when studying mobility data.

Acknowledgements

I'd like to thank all the people that were involved in the realisation of this thesis. Particularly my academic supervisor, Professor John Lee, how gave me precious help and advices throughout this important work.

I also could not have done this thesis without the amazing opportunity of an internship at the *Luxembourg Institute of Science and Technology*. I want to thank Fintan McGee who provided me guidance in my research and helped me formulate ideas. I'm thankful to all the people I met at *List* including Benoit Otjacques, Nicolas Médoc, Adrien Buche, and many other colleagues.

This work was done in collaboration with *Motion-S* and I want to thank the people there as well, especially German Castignani who gave me help and motivation for this thesis.

Finally, I'd like to thank my family for their support and patience throughout all my academic years at UCLouvain, and the help they provided for the writing of this paper.

Contents

1	Introduction	1
1.1	Motivations	2
1.2	State of the Art	3
1.3	Thesis structure	4
2	Data Analysis	5
2.1	Description of the datasets	5
2.2	Data Exploration	7
2.3	Pre-Processing	11
2.4	Additionnal Data	13
2.5	Analysis	16
2.5.1	Impact of features on the severity of an accident	16
2.5.2	Feature Importance	21
2.6	Discussion	27
3	Predictions	30
3.1	The models	30
3.1.1	Support Vector Machines	31
3.1.2	Platt's scaling	34
3.2	Training & Testing	35
3.3	Results	36
3.4	Improvements	39
4	Data Visualization	40
4.1	Theory	40
4.1.1	Definition	40
4.1.2	Motivations	41
4.1.3	Techniques	41
4.2	ICoVeR	43

4.2.1	Overview	43
4.2.2	Improvements	44
4.3	Other Visualizations	45
4.3.1	Maps	46
5	Conclusion	49
	Bibliography	51

List of Figures

1.1	Number of cars around the world between 1994 and 2013	1
2.1	Road representation by <i>Here.com</i>	6
2.2	Curvature of a line <i>@wikipedia</i>	9
2.3	Slope of a line	9
2.4	IRI roughness scale (Sayers et al., 1986)	10
2.5	SIN-COS encoding of cyclic feature	12
2.6	Linear Discriminant Analysis	14
2.7	Amount and percentage in accidents of various severity based on the day of the week	17
2.8	Amount and percentage in accidents of various severity based on month of the year	17
2.9	Amount and percentage in accidents of various severity based on the hour	18
2.10	Amount and percentage in accidents of various severity based on the speed limit	18
2.11	Amount and percentage in accidents of various severity based on the type of road	19
2.12	Amount and percentage in accidents of various severity based on their cause	19
2.13	Amount and percentage in accidents of various severity based on the age category	20
2.14	Amount and percentage in accidents of various severity based on different road signalization	20
2.15	ROC curve and AUC	24
2.16	ROC curve and Precision-Recall curve	26
2.17	Feature Importance based on Information Gain	28
2.18	Feature Importance based on Weight	29
3.1	Support Vector Machines	32
3.2	Kernels in Support Vector Classifiers <i>@scikit-learn.org</i>	33

3.3	Different cost and gamma on a SVM classifier with a RBF kernel @scikit-learn.org	34
3.4	Using Platt's scaling on SVM result [33]	35
3.5	Practical use of danger metrics	38
4.1	Types of dataset and its data @Tamara Munzner	42
4.2	ICoVeR at https://git.list.lu/eScience/ICoVeR	43
4.3	Added functionalities for ICoVeR	44
4.4	Zooms in interactive map of accidents	46
4.5	Accidents and their severity across cities in Luxembourg.	47

List of Tables

2.1	Classifier's scores on train and test set	27
3.1	Training of SVM using a polynomial kernel with various degree and cost	35
3.2	Training of SVM using a radial kernel with various gamma and cost . . .	36
3.3	Confusion Matrix SVM on test set	36
3.4	Classification score with SVM	36
3.5	probability distribution over classes per observation	37
3.6	Average probabilities in 'predictions' versus 'actual'	37

Chapter 1

Introduction

Since the beginning of the XX^{th} century, automobiles have been a significant part of the life of many people around the globe.

Used everyday by millions to go to work, carry goods, travel, or even for the pleasure or sports, more and more cars are getting introduced in the market[1].

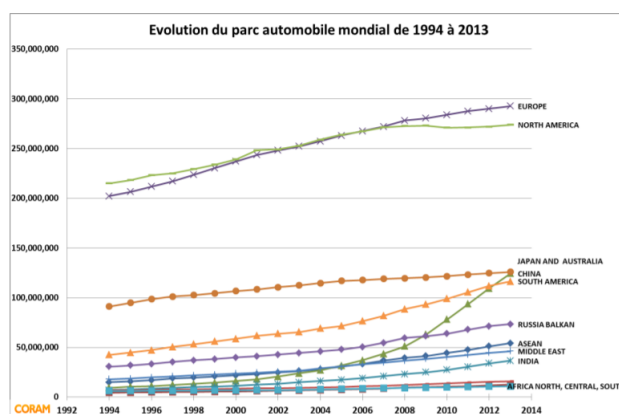


Figure 1.1: Number of cars around the world between 1994 and 2013

To allow this increase of automobiles, countries have invested billions in road infrastructures, specified rules to follow while driving, and defined regulations for car manufacturers to introduce new vehicles on the market. All to achieve the goals of **efficiency** and foremost **safety**.

Indeed, road safety has been a main issue in every country given that road transport is the most widely used means of travel and a primary cause of accidents[2].

This is why car makers are constantly innovating in making their vehicles safer, starting with the compulsory seat belts in 1968 and the introduction of airbags in the 1980s[3].

Governments are as well pushing for new measures to reduce accidents such as lower speed limits, stricter examinations for new drivers, police controls, etc.

Both are using previous accidents to learn and improve road safety by means of statistical analysis. With the evolution of Machine Learning and the arising capabilities of fast computing, new kind of algorithms and methods emerge to tackle these information related problems.

1.1 Motivations

Road traffic injuries constitute a major public health and development crisis, and are predicted to increase if road safety is not addressed adequately[4]. The use of modern statistical techniques, data mining and machine learning methods is the solution to find ways of improvements.

In cooperation with the *Luxembourg Institute of Science and Technology*[LIST] and *Motion-S*, we wanted to investigate what are the main factors affecting the severity of an accident using classifier algorithms such as Decision Trees; be able to formulate new metrics of 'danger' for any roads given previous crash history using Support Vector Machines; and explore ways to visually represent the data and how such representation can be useful for different actors affected by road safety.

There are multiple advantages in finding features affecting the most the severity of accidents. Drivers can instinctively identify segments of road that require more caution and a maximum of their attention. Governments can reduce the risk of accident by creating constraints for slower driving (speed cameras, police officers, speed bump, etc), placing regulatory signs to warn users of danger or to allow a safer traffic flow, or even improving the road infrastructure itself. Car makers can also benefit from such information to prepare their vehicles for real life road conditions, and in the context of self-driving cars, adapt the driving parameters based on the type of road taken.

Computing metrics of accident severity on roads can be useful to measure the 'risk' taken by a driver if he chooses to take a certain itinerary instead of another. *Motion-S* is looking for such measure of *spatial risk* to accompany their existing *behavioral risk* that any drivers is exposed to when driving.

Finally, finding ways to visualize the data will be useful to present our results and give the opportunity to any actor to explore different behaviors in traffic accident data.

1.2 State of the Art

Accurate estimations for the reasons of accidents are hard to make. Many different factors can be involved before the moment of impact and most of the time, only posterior observations are available. Nonetheless, many researchers have studied ways to understand causes and consequences of these accidents.

It can be simple statistics on what is the age range of drivers involved in most accidents or if they are more likely to happen on highways than on rural roads, etc.

More advanced analysis include the use of data mining techniques to identify locations with a high frequency of accidents and then analyze these locations to identify common factors that have an effect on road accidents. Starting by dividing the accident locations into k groups using K-means clustering algorithms based on road accident frequency counts. Then applying an association rule mining algorithm to highlight the relationships between distinct attributes which are in the accident data set[5].

Other ways of studying road accidents are to train classifier models such as Artificial Neural Networks[6], Decision Trees[7], and Support Vector Machines[8] which were reported as providing satisfying results.

In summary, literature on crash injury severity showed that significant attention has been focusing on crash severity modeling but injury outcome prediction was not a common major focus. Statistical models were more frequently used in crash severity modeling compared to machine learning methods, while machine learning methods were mostly used as prediction tools.

1.3 Thesis structure

All along this paper, theoretical aspects of the concepts used while manipulating data will be explained. This is an effort to maximize the comprehension of any readers, even if they are less experienced with Machine Learning or Data Science in general.

Starting from Chapter 2, the available data will be introduced along with a study on its quality and the early information it can give to the reader with the help of statistical tools and Decision Trees.

In chapter 3, a predictive model will be constructed using Support Vector Machines in order to estimates metrics assessing the danger that a driver can be exposed to on each road he takes.

Then, in Chapter 4, an introduction to the principles of Data Visualization, applied to our data with its advantages and the extra insights of information it can provide.

Finally, Chapter 5 will summarize the results gathered in this thesis and introduce possible improvements and additional leads to capitalize on the present work.

Chapter 2

Data Analysis

In this Chapter, we will discuss the data that we were provided with. We will then perform Data Exploration which is a capital step in every Data Science and Machine Learning project. Understanding the dataset that we have will allow us to identify potential outliers, help us perform pre-processing of the data to assure that the information we use is consistent and well formatted, and help us highlight useful information about traffic accidents. It is also important to find key characteristics about our dataset and if some features are related in anyway between them in order to choose an appropriate model and better tune its parameters. Once we are satisfied with the quality and the understanding of our dataset, we will try to identify factors that affect the severity of traffic accidents.

2.1 Description of the datasets

We were provided with two datasets.

- A large multivariate dataset containing data acquired by drivers in Luxembourg gathered by Motion-S. This database stores lists of trips performed by drivers in Luxembourg and in its nearest neighbour countries. Each trip is a sequence of geographical points taken roughly every meter gathering information about time, speed, acceleration, bearing, and the characteristics of the road. Characteristics considered are the speed limit, the curvature, the slope, the regulatory signs if any, and much more. This information is provided by the location and mapping company *Here Technologies*.
- A small dataset containing an historical overview of the road accidents that happened in Luxembourg between 2013 and 2017[9]. This set contains 4590 traffic accidents with their associated severity: if it was a *slight*, *serious* or *fatal* accident. Information

available for each accident are its geographical coordinates, the date and time, the age category of the drivers and an observation of the accident such as: against a tree, against a moving vehicle, ...

The main goal of this paper is to find ways to modelize traffic accidents in order to understand what are the dangers found on roads going from point A to point B. To do so, Motion-S was capable of mapping each accident's pair of coordinates to its corresponding road and retrieve the road characteristics with the services provided by *Here Technologies*[10], the same services they use to acquire road information in the first dataset.

Here.com is a geomapping service offering large databases concerning navigation, road infrastructure and maps. They record every road in the world and divide them into links. Figure 2.1 shows how each road is separated into links and each link is a segment of uninterrupted road. A new link start at each traffic sign encountered or at an intersection. With each link is associated metadata about the road such as the speed limit, the name, the geographical context, road quality, etc. It also contains a set of geographical points spreading along the link. Each point contains information such as the curvature of the road at that coordinate, the metadata associated to the link it is part of, the current slope in regard to the first point in this link, etc.

These are the points that Motion-S helped to match with our historical accident records to give us the dataset with all the accidents, their severity, and the road characteristics associated with each accidents. All this rich information will be the focus of this paper.



Figure 2.1: Road representation by *Here.com*

2.2 Data Exploration

Let's begin by identifying 3 general characteristics proper to every dataset[11]:

Dimensionality What is the number of attributes in our dataset? We start with 82, but this number will evolve during the whole analysis with Dimension Reduction and Enrichment. But it is worth noting that we want to avoid having too many attributes in order to stay away from the *Curse of Dimensionality* where the higher the dimension of our set gets, the more the observations become sparse; and the more sparse our data is, the less statistical significance our results will have.

Sparsity Is data present in the majority of observations or on the contrary do we have a lot of missing information or values set to 0? In our case, most of the features contain values and the majority of missing values concern a specific feature that can be ignored.

Resolution Have we enough observations to find patterns in accidents? There was "only" 4590 accidents in Luxembourg between 2013 and 2017. In a way it is positive to have a low number if it concerns traffic accident. However, for Machine Learning models, it is always better to have more observations available but we will do with what we have!

It is never a good idea to assume that the data is perfect. There can be incorrectness in data due to multiple reasons such as human error, limitation in measuring devices, flawed data collection,...

So before continuing, let's acknowledge the flaws in our dataset:

Measurement Error The measurements of the accidents are not guaranteed to be accurate and are done by humans. The place of impact can be meters away from the coordinate registered, and make us blame another part of the road.

Data Collection Error We are missing important information, for privacy reason (such as age, sex) or that were deemed unimportant, unnecessary for the simple task of counting accidents (such as type of vehicle), or even are just impossible to collect (speed during the crash)

Noise There is noise in our data. Road information is retrieved from a single static point. We cannot guarantee a precise measure of every feature.

Outliers	There are outliers in our data. We will see later when we try to visualize the data that some records stand significantly apart from the rest due to unreasonable feature's values in some observations.
Class Balanced	Since we are dealing with a labeled dataset (the severity of accident), we can assess if the classes are well balanced or if on the contrary, one class is over represented. In our dataset, we find 3247 slight, 1188 serious and 155 fatal accident, respectively 71%, 26% and 3%. So it is indeed hugely imbalanced in favor of <i>slight</i> accidents. Meaning it will be difficult to isolate the factors affecting most <i>fatal</i> accident. But we can nonetheless identify in which situation a <i>serious</i> or <i>fatal</i> accidents is more likely.
Up-to-Date Data	Results given by <i>Here.com</i> are the most up-to-date available. Road information added to each observation of accident were requested in March 2019 and reflect the state of these roads at this period. A road that witnessed a crash in 2013 might have since then been redone or changed (new pavements, change of speed limits, etc). Because we cannot retrieve the state of the road at the time of the accident, we will assume in the following analysis that the roads have not been significantly changed in recent years.

Variables Each observation carries a set of features describing the context of the accident and the characteristics of the road[12]. Some of the main ones are:

Date (*yyyy:MM:dd:hh:mm*).

Severity (*Slight, Serious or Fatal*).

Observation (*Degree*).

Age Range (*9 Categories*).

Latitude (*Degree*).

Longitude (*Degree*).

Curvature ($10^{-6}/m$).

Slope (*Degree*).

Speed Limit (*7 Categories*).

Urban (Yes/No).

Signs (List of signs present).

Iri (m/km).

It is important to understand each variable available, the unit it is expressed in, and the information it carries. While the meaning of most of them is quite clear, it is worth giving a short explanation for a couple of them.

Curvature[13] We are given features *Radius* and *Curvature*.

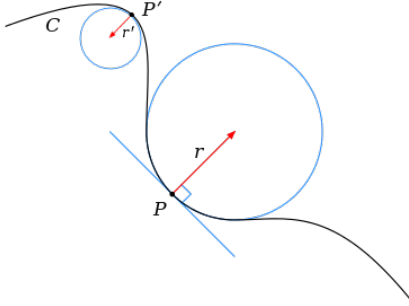


Figure 2.2: Curvature of a line
@wikipedia

To compute the curvature at point P, we imagine a circle that best approximates the curve at that point. Then with the radius of this circle, we simply compute $Curvature = \frac{1}{Radius}$. The smaller the radius is, the higher the curvature will be and thus sharper will be the turn at point P. The radius is also available in the data, allowing us to verify the validity of the curvature reported and complete the observations where it was set to

NaN.

Since both variables are directly linked to one another, we will only keep the curvature in our models.

Slope

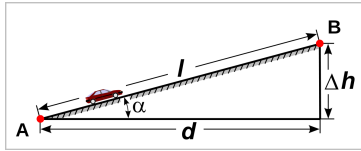


Figure 2.3: Slope of a line

Since *Here* considers a road as a sequence of segment of various size, the slope at point B is computed based on the first point A of the segment it is part of. In our dataset, the slope is expressed in 3 manners: Unit, percent and degree.

$$Unit = \frac{\Delta h}{d}$$

$$Percent = \frac{\Delta h}{d} * 100(\%)$$

$$Degree = \alpha(^{\circ})$$

Again, there is no information gain in keeping the 3 features that express the same value. We will first use them to verify the validity of *percent* and only keep the later in future development since *percent* is the unit commonly used on roads to inform about slopes.

Iri[14] Stands for "*International Roughness Index*" and is a measure of irregularities in the pavement surface that adversely affect the ride quality of a vehicle. We won't enter into details on its computation but it is worth remembering that the lower the Iri is, the smoother the surface of the road is.

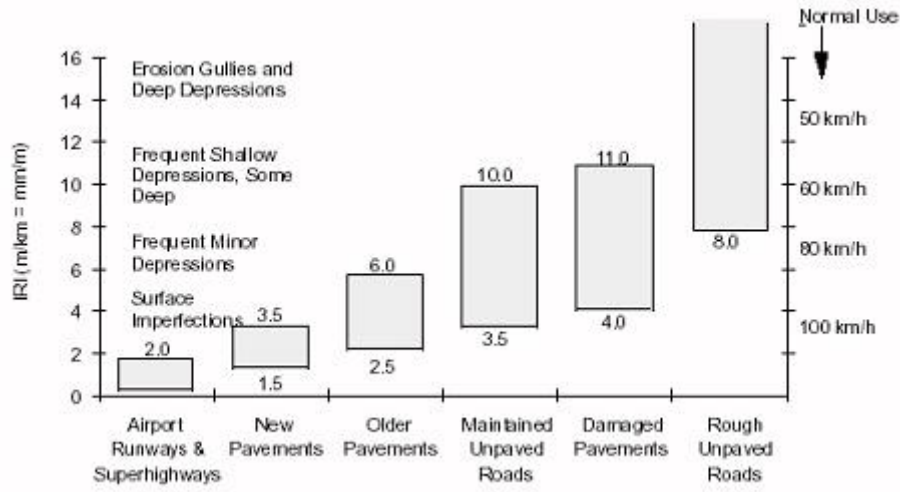


Figure 2.4: IRI roughness scale (Sayers et al., 1986)

2.3 Pre-Processing

Before going any further in the handling of our dataset, we need to ensure that we can "trust" its content. There might be missing variables or values that don't make sense in some observations and this would have an impact on the result of any analysis that we perform.

The crucial step of "Pre-processing", strongly advised in any Machine Learning project, is to perform data-cleaning, scaling continuous values within a specific range (*normalisation*), and potentially perform dimension reduction to lower the amount of variables.

Data-Cleaning In this step, we identify rows that contain empty values and remove the feature for all observations if it is missing for the majority of them. Sometimes however, the missing values are scattered across features and observations but not in a significant amount. In this case, we can either ignore them while building a model or we can try to estimate them. The later option is preferable since the dataset is already small. To do so we can take the average of the same feature across all observations, or compute it based on other features. The curvature for example can be computed from the radius as explained earlier, and there are multiple units of slopes available to help us compute the percent.

In parallel, we need to deal with values that don't make sense such as a speed limit of 9999km/h or a slope of 1000°... These anomalies can result from an error of encoding or are just another default value like NA (*"Not Applicable"*). If the curvature and slope are expressed in unreasonable values, we consider the segment of road as straight and flat. If the speed limit is over the maximum possible in the country (130km/h), we will look at what is the average traffic speed on this road and set the speed limit to the closest category (30, 50, 70, 90, 110, 130 km/h).

Normalisation Next, we need to ensure that all continuous features are scaled on the same range to avoid comparing Apples and Oranges. To do so, we use the Min-Max Normalisation that consists in finding the spread of the feature $[min - max]$ and report it to a scaling $[0 - 1]$.

Other features that should be subject to normalisation are cyclical ones such as time. Indeed we currently have resp. day of week, month, and time but they are represented as numerical values ranging for resp. $[1 - 7]$, $[1 - 12]$, $[0 - 24]$. However hour 23 is closer to hour 0 than it is to hour 21 and this nuance is missing currently. To solve this issue we will use the sin-cos encoding[15]:

Let's imagine a list with every hour ranging from 0 to 23. Like in the normalisation, we

will change the range but instead of $[0 - 1]$, we will set it to $[0 - 2\pi]$. Then compute the sinus of each hour. Now, the information carries that 23:00(-0.26) is closer to 00:00(0.0) than to 21:00(-0.71), but it is also close to 12:00(0.0)! To avoid this, we add another feature which is the result of the cosinus of each hour. This way, 23:00(0.97) stays closer to 00:00(1) than to 21:00(0.71) and far away from 12:00(-1). Sin and Cos considered at the same time gives a unique cyclic representation of all the hours in a day.

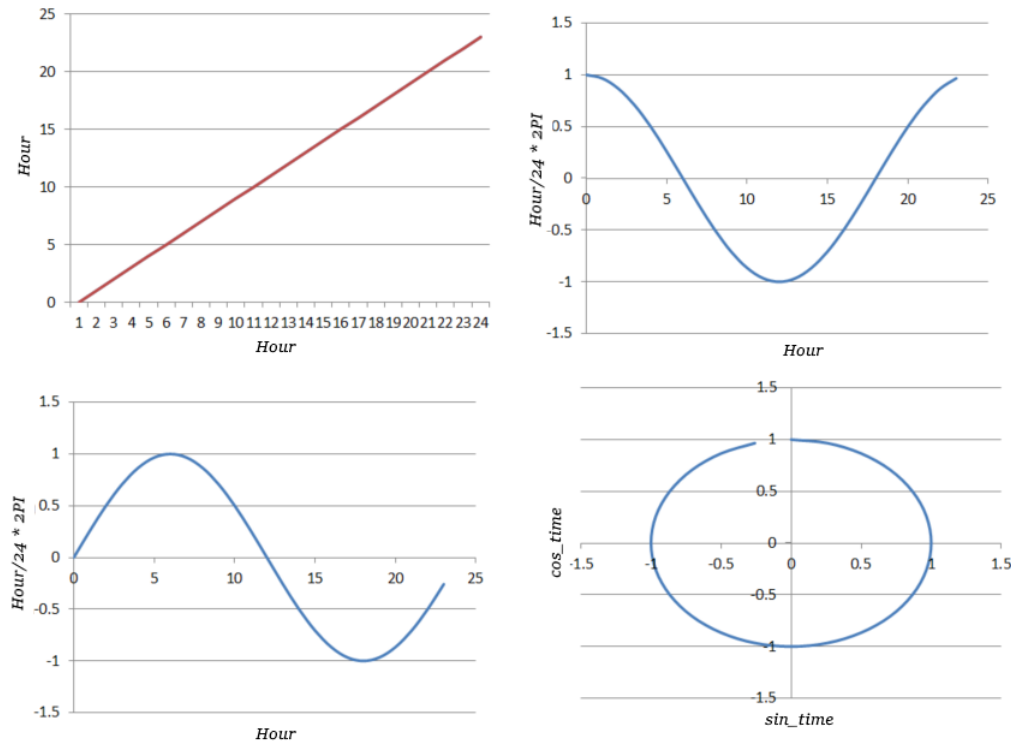


Figure 2.5: SIN-COS encoding of cyclic feature

Dimension Reduction Finally, this step consists in removing or summarizing features in our dataset in order to reduce the load of input while building a model; and to help us having a visualization or a representation of our data. It is indeed easier to visualize a 3-Dimensional dataset than a 25-Dimensional set. This is usually performed by statistical techniques such as Factor-Analysis or Principal Component Analysis[PCA] but they provided poor result on our dataset so we will only remove bad and "replicated" features for this task.

There is however a method closely related to PCA better adapted to our dataset[16], *Linear Discriminant Analysis* [LDA][17]. Both techniques look for linear combinations of variables which best explain the data but LDA explicitly attempts to model the difference

between the classes while PCA does not take it into account.

In our mutli-class dataset, consisting of a set of sample $x_1, \dots, x_n$ with labels $y_1, \dots, y_n$, LDA will be based on the analysis of two scatter matrices:

within-class $S_w = \sum_{i=1}^n (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$ where μ_k is the sample mean of the k -th class.

between-class $S_b = \sum_{k=1}^m n_k (\mu_k - \mu)(\mu_k - \mu)^T$ where m is the number of classes, μ is the overall sample mean, and n_k is the number of sample in the k -th class.

To find the set of linear combinations (with coefficients $\mathbf{w}$), we need to solve the optimization problem that maximizes the ratio of the *between-class* scattering to the *within-class* scattering:

$$\hat{w} = \operatorname{argmax}_w \frac{w^T S_b w}{w^T S_w w}$$

The solution is given by the generalized eigen value problem $S_b w = \lambda S_w w$

A working implementation of LDA is available with the python library `scikit-learn`¹. It allows us to have a general representation of our data as in Figure 2.6. This gives us the clear message that even on the two principal axes computed by LDA supposed to represent and separate the dataset the most, we have a clear overlapping of our classes. *Slight* accidents are more represented and more spread across the axes. It makes sense since some observations of *slight* accidents cannot be reproduced to have a *fatal* one (going 30km/h is unlikely dangerous). But any *deadly* accident could have been less serious depending on unavailable variables (speed before the crash for example) or just a better luck for drivers involved. We also notice outliers that stick out that we will be able to understand in a later chapter.

2.4 Additionnal Data

One of the problem highlighted before about traffic accidents or mobility datasets in general is that a lot of factors can affect driving. In our dataset, we are lacking the type of vehicule, the speed during the impact, and all exterior factors such as weather, a potential blinding by the sun, traffic, eventual distraction along the road, ...

¹https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminant

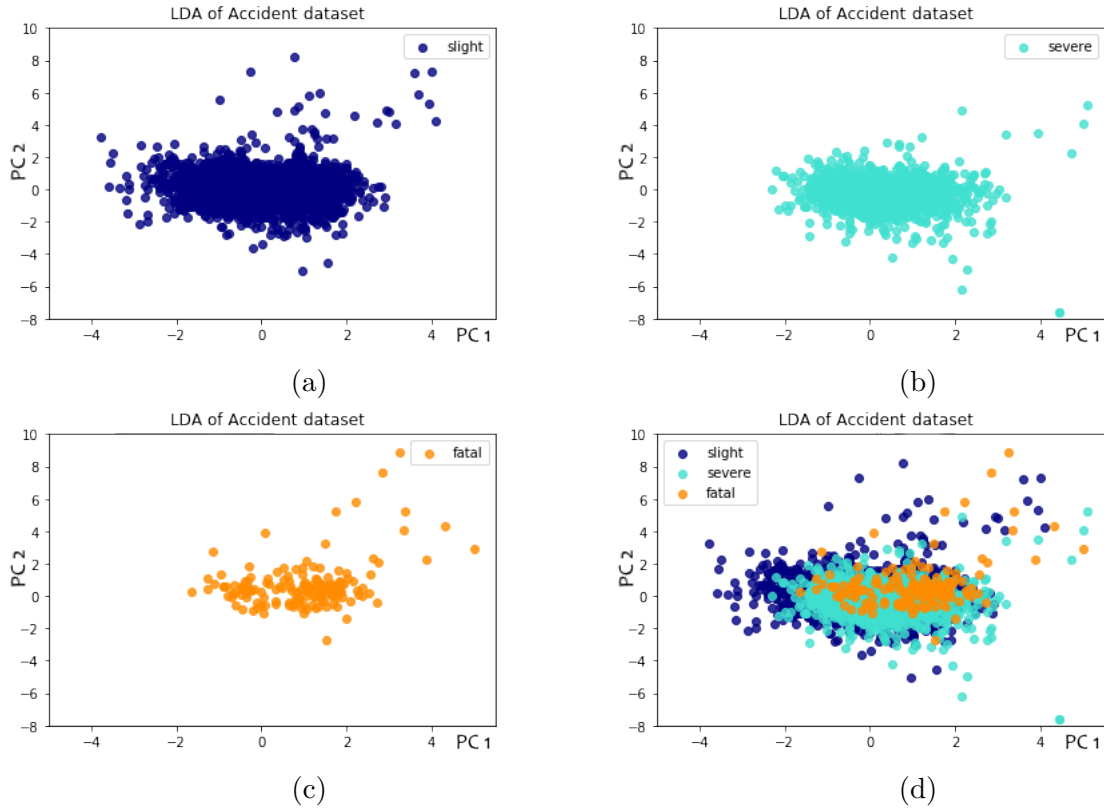


Figure 2.6: Linear Discriminant Analysis

A lot of information cannot be retrieved for privacy reason or just by absence of measurement tools. However it is worth trying to enrich our dataset with all the information available that could have an impact on traffic accident.

Weather Data Using the API *worldweatheronline*², it is possible to retrieve the weather forecast of the closest meteorological station to a pair of GPS coordinates at a certain date. The additional information provided are the day's temperature (max and min), the precipitations, wind, visibility, cloud cover, snow, etc. It also provides the time of sunset and sunrise of the day, allowing us to assess if an accident occurred at night or at daylight.

Traffic Flow *HERE.com* provides a traffic speed pattern per segment of road. In our dataset, two features `trafficTo` and `trafficFrom` are present. They give information about the average traffic speed in both direction (or one if it is a one way street) every day. If not empty, they give a list of 7 numbers or indexes, one per day of week, starting from Sunday. One index corresponds to an entry in an external dataset containing all traffic pattern possible.

²<https://www.worldweatheronline.com/developer/api/>

A pattern is a sequence of speed observed, with one observation every 15 minutes. Thus there are 96 speed records per day.

This information can be used to assess the *popularity* of a road or give a measure of *congestion* at a certain time of day.

First, to measure the popularity of a road, we can look at the 7 patterns and observe if the traffic speed stays roughly the same throughout the week; or if in the contrary we notice significant drops in the traffic speed at certain hours or certain days. Our measure of *popularity* would then be the standard deviation in the sequence of speed during the whole **week**.

$$popularity = \sigma_{week} = \sqrt{\frac{\sum_{t=1}^{n=7*96} (x_t - \mu)^2}{n}}$$

where $[x_1, \dots, x_n]$ is the traffic pattern of a whole week and μ is the average speed during the week.

Since we are given a time for the accident, we can also look up what was the average speed observed at this moment and infer if there was a decrease of speed resulting in traffic at that time or not. To give a measure of this *congestion*, we take the standard deviation of speed during the whole **day** and compute how many 'steps' away we are from the speed limit.

$$\sigma_{day} = \sqrt{\frac{\sum_{t=1}^{n=96} (x_t - \mu)^2}{n}}$$

$$congestion = (SpeedLimit - Speed_t) / \sigma_{day}$$

Where $[x_1, \dots, x_n]$ is the traffic pattern during the day of week of the accident, μ is the average speed during the day, and $Speed_t$ is the traffic speed recorded at the hour of the accident.

Signalization We can find signalization signs in 30% of the observations in our dataset. Their value goes from traffic light, yield, ..., to warnings in icy conditions. The most observed ones are the "overtaking" signalization, "road profile"(sharp turn or intersection ahead, etc.), and "crossing". If we could have had road signalization in most record of accidents, it could have been a really useful information to provide context about regulation of the road and the awareness given to the driver. However, the signs are only reported if they are present in the segment of road considered by *Here.com* and not if they concern the whole road. For example, we might be looking at a segment of road that does not allow overtaking, but since the sign

announcing the rule is 100m behind, in another segment of road, it will not be reported by *Here.com* and we would wrongly consider the road as allowing overtaking.

Thus, considering this information for building models would be a bad idea since the absence of sign data doesn't mean there were no rules on this road, making it an unreliable feature. Nonetheless, as we will see in the following section, we can isolate the subset of our dataset that offers sign information and identify if some traffic rules influence more or less the severity of accidents.

2.5 Analysis

In this section, we will start by retrieving and picture simple statistics about the severity of accidents based on independent features. We will then move on to introducing Decision Trees and their ability to highlight the *importance* of features. And finally, we will conclude with a small discussions on our findings.

2.5.1 Impact of features on the severity of an accident

Let's find out the eventual impacts on severity that are linked to some categorical features. We will start by retrieving the amount of accident per severity for each categories within a feature. Then put it in perspective by computing the ratio of accidents in regard to the total within each category.

Note that we lack the likelihood of each category in a feature. We know that each day of the week will have the same probability of happening but we don't know the proportion of daily traffic on highways compared to the daily traffic on urban roads in Luxembourg. If we notice 4 times more accidents in the later, it is maybe because there are 4 times more urban roads than highways. However, if this number is actually 10 times more, then that would imply that it is more frequent to have an accident on highways. But without such information, we will refrain from jumping into conclusion too quickly.

Day of week

Figure 2.7: We don't really notice a different behavior in accidents during the first five days of the week. However we can see a decrease in the number of accidents during the weekend but with an increase in the proportion in serious and fatal accidents in these two days.

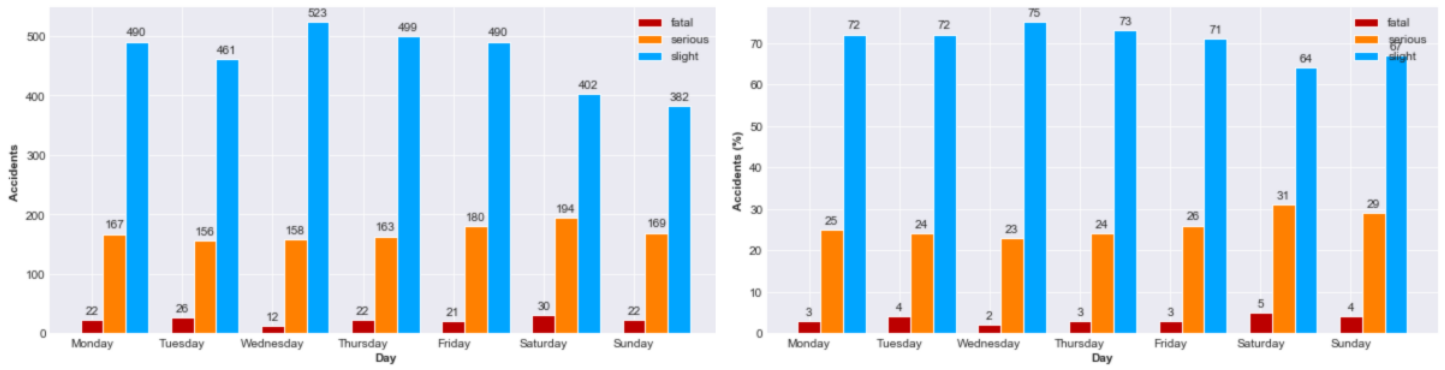


Figure 2.7: Amount and percentage in accidents of various severity based on the day of the week

Month of year

Figure 2.8: Even though winter time in Luxembourg often brings bad conditions for driving, we see a slightly higher likelihood of more serious accidents during spring and summer time.

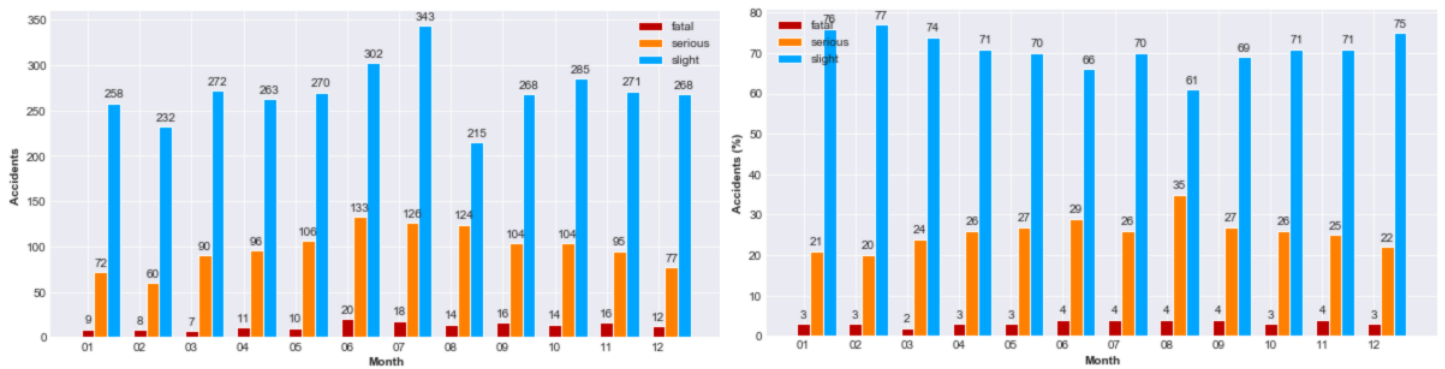


Figure 2.8: Amount and percentage in accidents of various severity based on month of the year

Hour of the day

Figure 2.9: It comes as no surprise that more accidents occur when more people are on the road, i.e. during rush hour. Although it is worth noting that during these hours, it results more often in *slight* accidents and that the more serious severity are more represented at other times, especially at night.

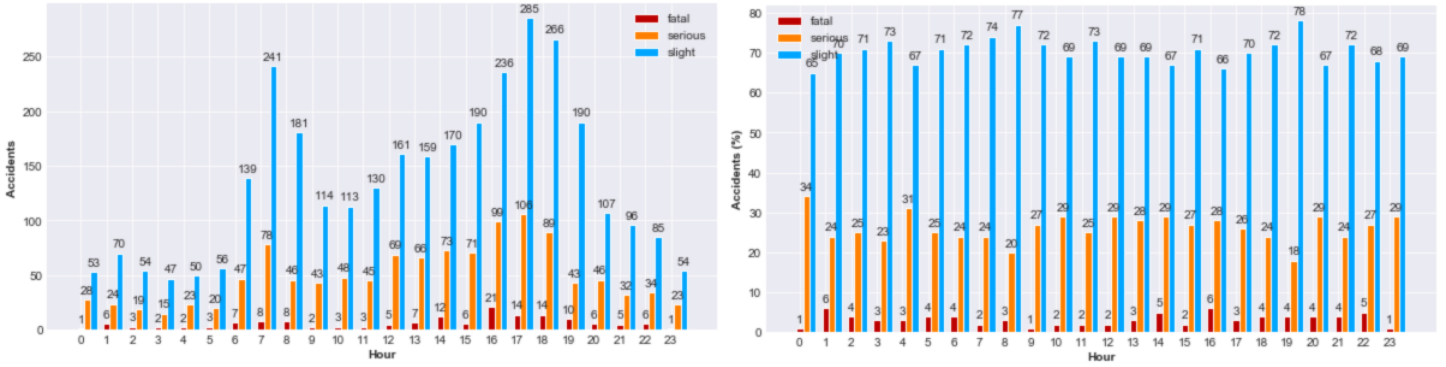


Figure 2.9: Amount and percentage in accidents of various severity based on the hour

Speed Limit

Figure 2.10: We can understand here that most accidents have happened in urban areas where the speed limit is set at 50km/h. Most of these accidents were of *slight* severity and it is understandable. Low speed prevents having severe crashes, urban areas are places where traffic tends to concentrate, resulting in more occasions for accidents. The higher the speed, the more likely the severity will be higher.

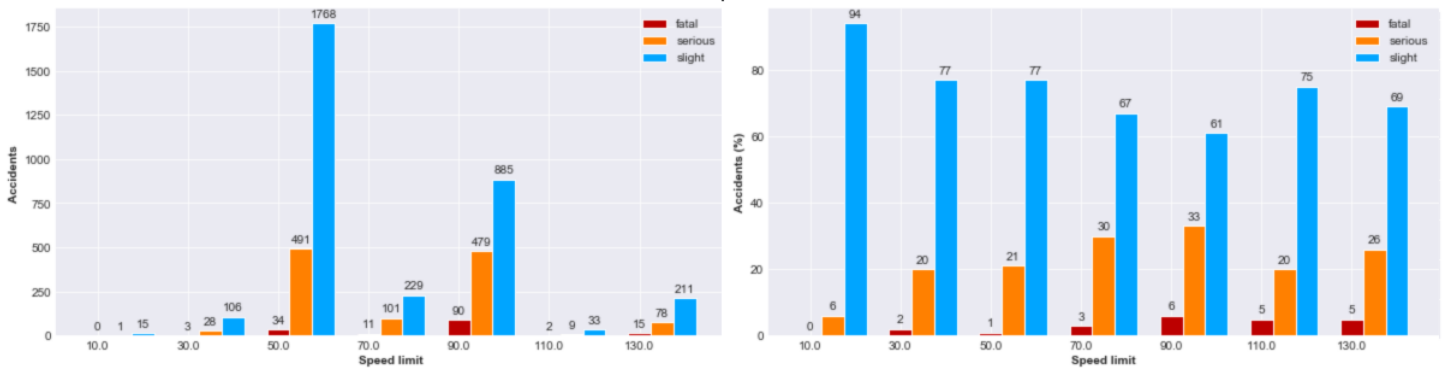


Figure 2.10: Amount and percentage in accidents of various severity based on the speed limit

Road Type

Figure 2.11: And this confirms the intuitions above that roads allowing higher speed limit observe fewer accident but with higher likelihood of serious accidents.

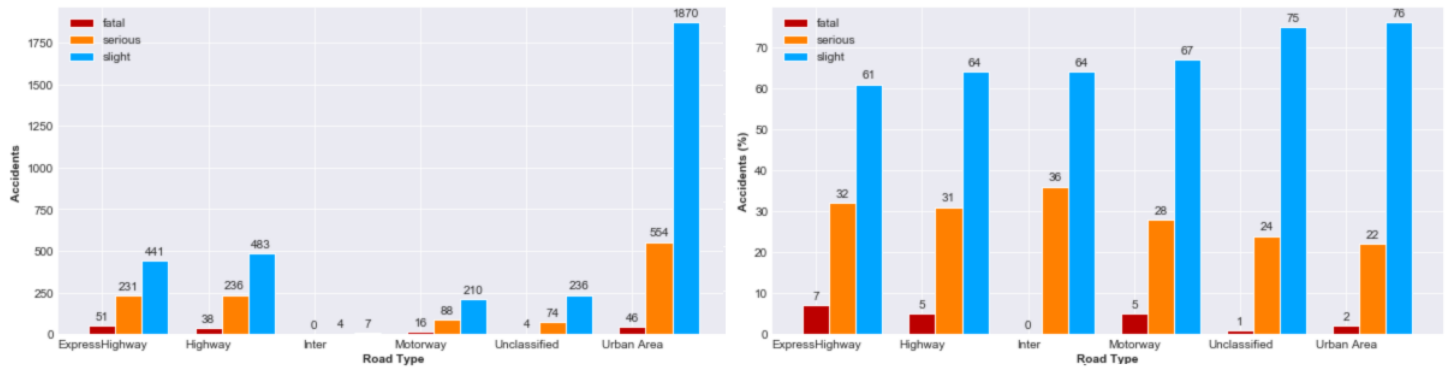


Figure 2.11: Amount and percentage in accidents of various severity based on the type of road

Cause of the accident

Figure 2.12: Even though we are mainly interested in what are the road characteristics impacting the most the severity of an accident, it is worth looking into what context accidents more often result in more serious outcomes. We can highlight that important risks are fixed obstacles and particularly trees. We notice a rise in more serious accidents when such observations were reported.

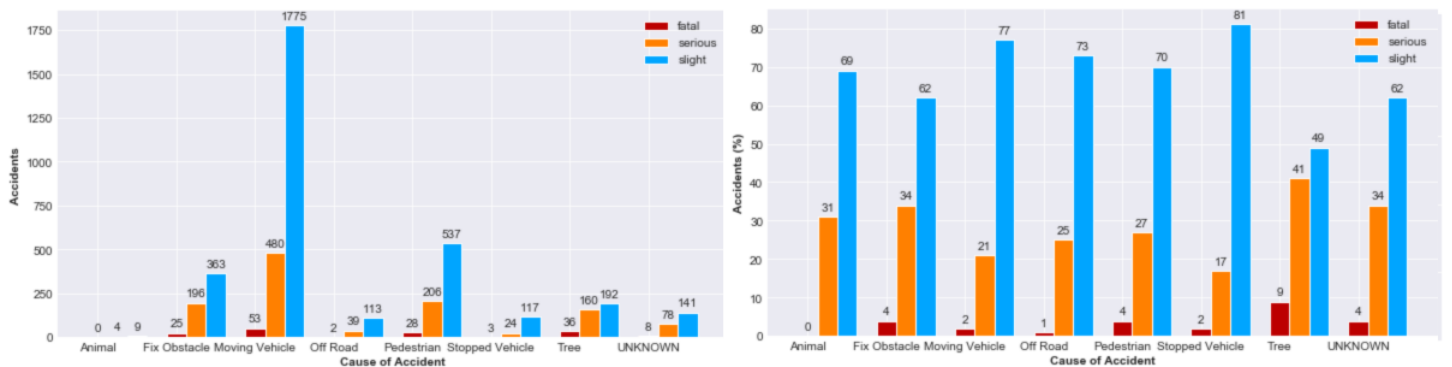


Figure 2.12: Amount and percentage in accidents of various severity based on their cause

Age of drivers

Figure 2.13: With the age feature, we can see that most accidents involve younger drivers but the risk of a more serious accident increase as the drivers get older.

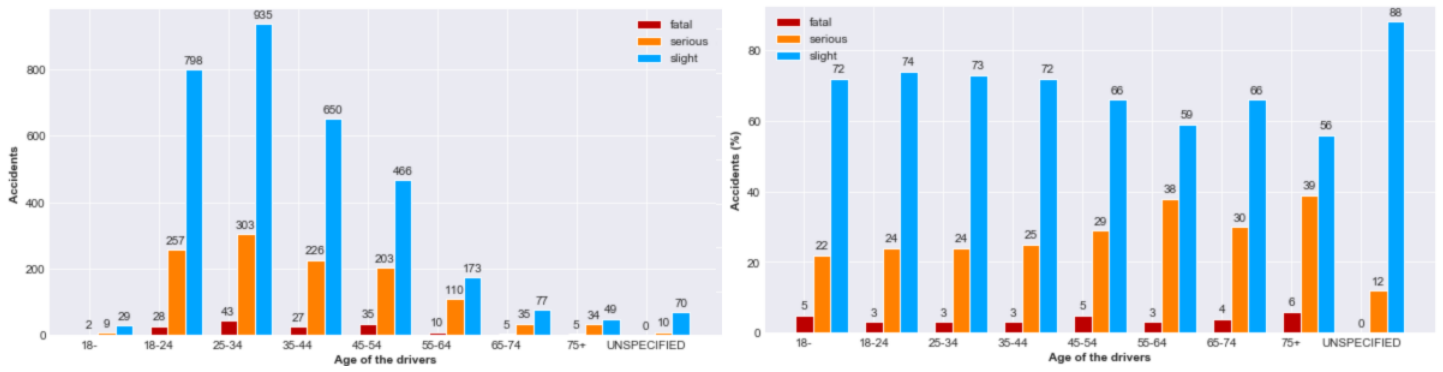


Figure 2.13: Amount and percentage in accidents of various severity based on the age category

Road Signs

We finish by taking a quick look at what insights the road signs can provide to us. As mentioned before, only 30% of accidents report a sign nearby but that doesn't mean that the other 70% don't have rules or caution that apply. The only comparison we can make is by taking the proportion of accident of different severity to the overall proportion of all accidents with sign information.

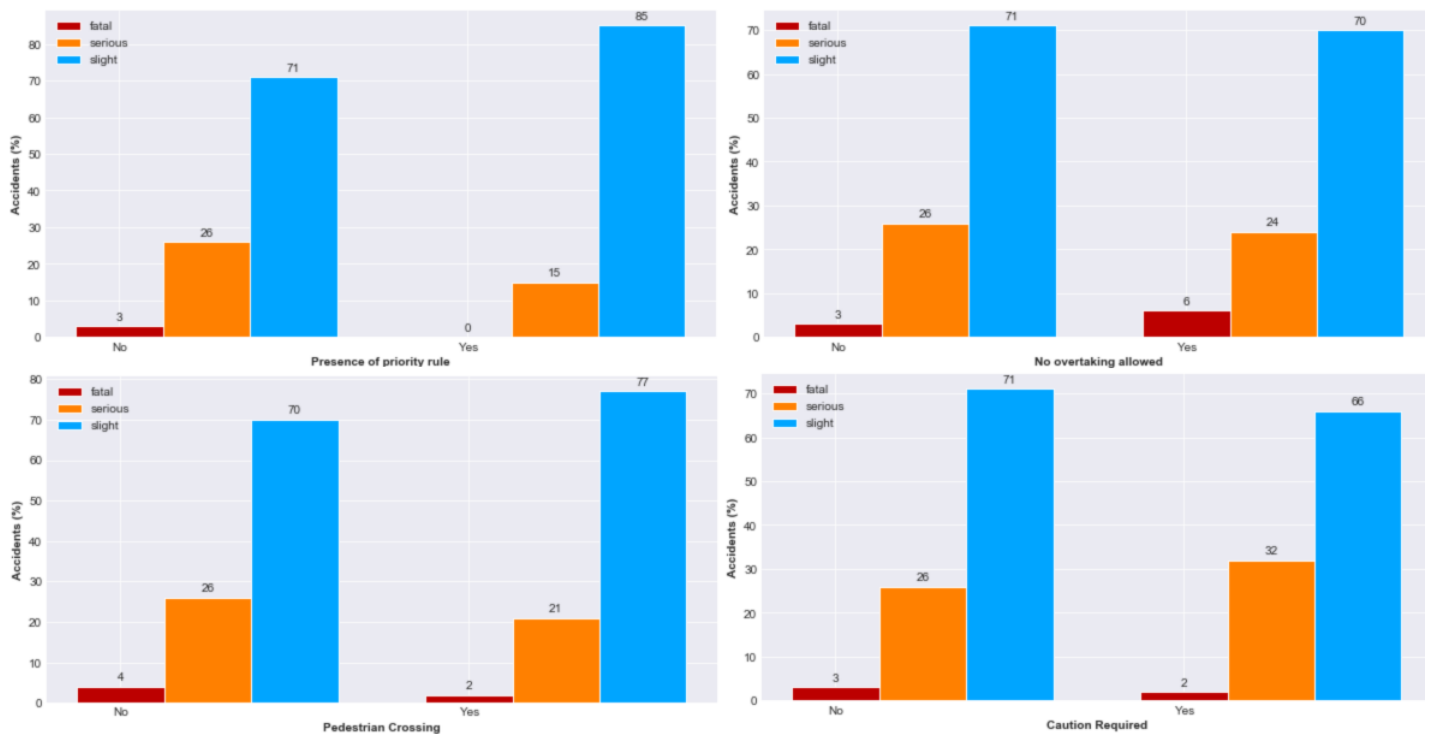


Figure 2.14: Amount and percentage in accidents of various severity based on different road signalization

What we can remember from this four results is that on roads with signs forcing us to slow down due to a traffic light, stop, yield or pedestrian crossing, we observe a higher portion of slighter accidents. However, on roads that warn drivers of dangers with "caution" or "no overtaking" signs, we notice a higher chance of having a serious or fatal accident.

2.5.2 Feature Importance

An important aspect of our dataset is that it contains a lot of categorical features (age-range, speed limit, observations, road type, etc) and it is labelled. Those characteristics make our data well suited for *Decision Tree* models[18].

Decision Trees are simple yet powerful supervised algorithms in Machine Learning. They provide one of the fastest ways to identify most significant variables in relation between two or more variables[19]. Indeed, to build a tree in order to classify a dataset, we rely on techniques to find the best features to separate our target class[20]. Let's imagine that 80% of fatal accidents happen at night time, this information would be one of the first things we look at when we want to predict the severity of a new accident. The most popular of these techniques are:

Gini index It is calculated by subtracting the sum of the squared probabilities of each class from one for each feature. We choose the root of the tree by the feature that has the least Gini index then compute the new Gini indexes for the remaining features on each branch of the root. Then again split the tree using the variable with the least index and so forth until the set is classified or until we met our end criteria. This approach favours larger partitions in splitting and is used in CART trees.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

where p_i is the probability of class i out of C, according to the current feature selected.

Info Gain It is retrieved using the Entropy. If a sample is completely homogeneous, it will have a entropy of 0. If the sample has an equal proportion of observations of each class, then the entropy is 1. The smaller the entropy is, the purest is the sample.

$$Entropy = -p \log_2(p) - q \log_2(q)$$

where p and q is the probability of success and failure respectively in that node. The gain is the expected reduction in entropy due to sorting on a chosen attribute and is obtained by computing the entropy of the parent node and

subtract to it the weighted average entropy of all sub-nodes available in the split.

$$Gain(S, x_j) = Entropy(S) - \sum_{v \in Values(x_j)} \frac{|S_v^j|}{|S|} Entropy(S_v^j)$$

where $S_v^j = \{\mathbf{x} \in S | x_j = v\}$.

It favors smaller partitions and is used in ID3 or C4.5 trees.

There are many others techniques such as *Chi-square*, *Reduction in variance*, etc but we are interested here in finding the importance that each feature brings while classifying our set and can use the information gain as a good metric.

Now that we've covered how we are going to build a tree, the next question is how do we assess that the tree is good enough to describe our dataset ?

A problem with decision trees is that they are prone in over-fitting. If our dataset is consistent, i.e all observations sharing the exact same features have the same target value, there will exist a tree that exactly describe our data, but only this one. We want to limit the size of the tree in order to generalize the classification but at the cost of accuracy in prediction.

To balance these two constraints, we will use not one, but a multitude of different decision trees, called a forest. Requesting a prediction from this forest will just be taking the majority of the votes from the trees it is composed of.

They are two main ways to build such forest:

Bagging[21] Decision trees are trained on randomly sampled subsets of the data, while sampling is done with replacement. This is the main logic behind Random Forests.

Boosting[22] Decision trees are trained on sampled subsets as well, but each sample is weighted so that those that were predicted incorrectly so far get a higher weight and are thus more represented in future sampling. The idea behind this is that difficult cases should be emphasized during learning compared to easy cases.

It is worth noting that bagging can be paralleled while boosting is only performed sequentially.

For this work, we will use gradient boosting provided by the python library `xgboost`³ and the rich library of machine learning techniques `scikit-learn`. We will also consider *serious* and *fatal* accidents as one class. Giving us a classification problem with only two target classes, easier to apply to decision trees and allowing us to re-balance a bit the

³<https://xgboost.readthedocs.io/en/latest/parameter.html>

dataset: 75% slight and 25% serious accidents.

The *gradient* in the name means that we will try to minimize a loss function, similarly to Neural Networks that use *gradient descent* to optimize, or *learn*, weights[23]. Once a tree is built, we compare its predictions to the actual outcome expected and the distance between prediction and truth represent the error rate of our model. We use these errors to calculate the gradient, which is the derivative of our loss function. For the following model, we will use the *Area under the curve*[AUC]

This curve is known as the *Receiver Operating Characteristic curve*[24] or [ROC]. It is a representation of the performance of a classification model at all classification thresholds. It relies on two parameters:

True Positive Rate Obtained by dividing the number of True Positives by the total number of actual positive observations, guessed right or wrong.

$$TPR = \frac{TP}{TP + FN}$$

False Positive rate obtained by dividing the number of False Positives by the total number of actual negative observations, guessed right or wrong.

$$FPR = \frac{FP}{TN + FP}$$

With **TP**:*True Positive*, 'the prediction was *serious* for an actual *serious* case', **FP**:*False Positive*, 'the prediction was *serious* for an actual *slight* case', **TN**:*True Negative*, 'the prediction was *slight* for an actual *slight* case', and **FN**:*False Negative*, 'the prediction was *slight* for an actual *serious* case'.

Let's illustrate this concept with the classification of *slight* or *serious* accidents based on a single parameter *speed* shown in Figure 2.15 with simplified values. Starting at the representation of the data in (a), we can fit a logistic regression curve to transform the *y* axis into the probability of an accident being serious. Since we need to classify each accident, we have define a threshold of probability value at which an accident is considered severe or not. Consider we pick the default 0.5, we get the confusion matrix:

		Actual	
		Slight	Serious
Predicted	Slight	5	2
	Serious	1	3

Giving us $TPR = \frac{3}{3+1} = 0.75$ and $FPR = \frac{1}{5+1} = 0.17$

Now, depending on what threshold we choose (c), we will get different classifications and so different confusion matrices.

The ROC curve represents these different confusion matrices given by different thresholds (d). This will allow us to identify the best threshold to classify our data, depending on our degree of acceptance of false positive. This curve also allows us to compute the AUC. If the AUC is 1 then our model perfectly classifies the data, if it is 0 then our predictions are always wrong. We will use the AUC to compare classifications: let's say a model based on *speed* gives an AUC of 0.8 and another model based on *curvature* gives an AUC of 0.65, we can say that *speed* better classify our data.

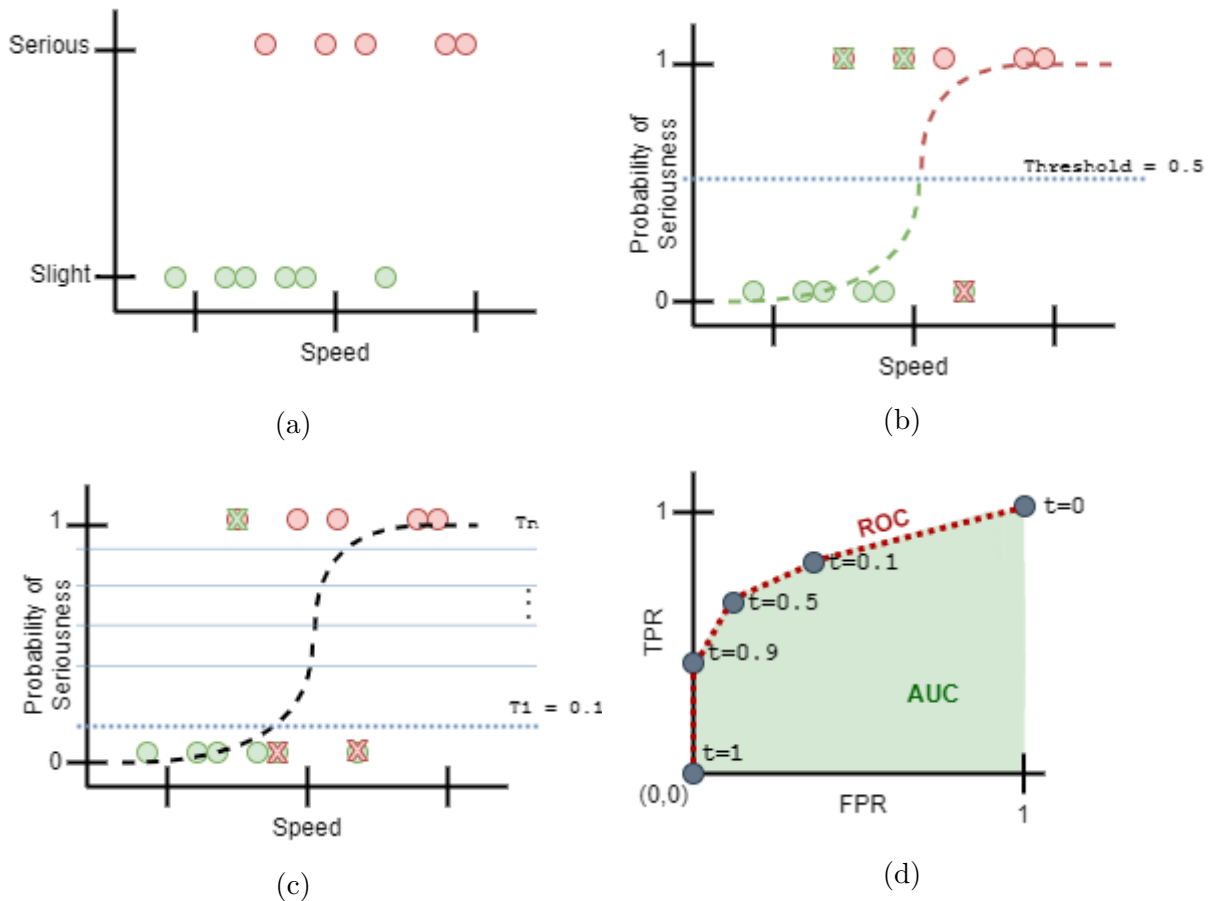


Figure 2.15: ROC curve and AUC

This is the logic that we will use in our gradient boosting model. Comparing the AUC of each new sub-tree and looking to maximize it. The creation of new trees stops if we see no improvements during 100 rounds or if we reach a limit of 300000 trees. We divide the dataset into a training set on which we compute our model and a test set, respectively 85% and 15%.

Now that our model is built, we will examine its quality, its performances, and what useful information it can bring us.

A first common thing to do is to measure the **accuracy** or what is the rate of correct predictions by the model. We achieve an accuracy of 67% in our test set but in our case, it is not the most valuable measure. Indeed, we could just predict *slight* at every observation and achieve an accuracy of 75% since the data is majorly composed of such observations. However, we mainly want a model that helps us understand what feature increase the most the likelihood of a *serious* accident. Ignoring this under-represented class to achieve high accuracy would not help at all. So we need to go further.

As previously discussed, ROC can give us a representation of how well a model is capable of classifying data. As shown in Figure 2.16 (a), the model is doing a great classification of the training set but is weak at classifying the test set. The AUC is 0.68 and the classifier nonetheless performs better than a random guessing. The threshold that can be defined to separate the probabilities between the two classes is 0.3.

But since our dataset is unbalanced, we could wonder how good the model is at correctly predicting serious accidents amid all the slight ones that tend to influence the final classification.

We use the **Precision** for this purpose. It helps us identify what proportion of *serious* severity predictions was actually correct. A precision of 1 would mean that we never declare an accident *serious* when it is in fact slight.

$$Precision = \frac{TP}{TP + FP}$$

Our model at threshold 0.3 has a precision of 0.42, or in other words, when our model predicts a serious accident, it is right only in 42% of the cases.

On the other hand, the **Recall** can help us identify what portion of actual serious accident was identified correctly. A recall of 1 would mean that no *serious* accidents were identified as *slight*.

$$Recall = \frac{TP}{TP + FN}$$

Our model at threshold 0.3 has a recall of 0.61, so has been able to identify 61% of the *serious* accidents.

We can interpret these two last measures as the model trying to compensate the class imbalance by assigning more easily *serious* severity to an accident, explaining the low precision since a lot of crashes are *slight*. But it then gives a more decent recall considering that it is hard to do the distinction between the vast majority of *slight* accidents and more *serious* ones.

Similarly to AUC, another way to train our model could be to maximize the *Area Under the Precision-Recall Curve*[25] [**AUPRC**] shown in Figure 2.16 (c), but it has provided poorer results in our case due to the difficulty of achieving high precision with the class imbalance. So we kept the maximization of AUC.

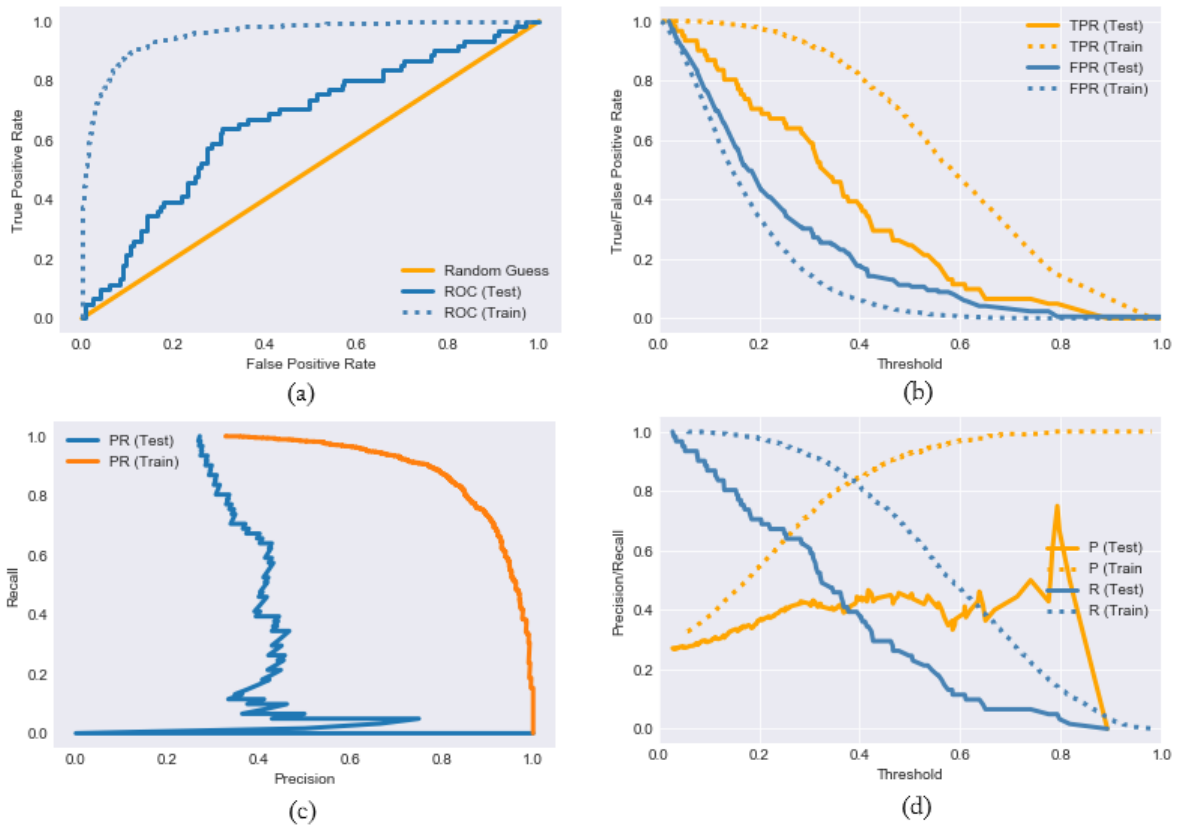


Figure 2.16: ROC curve and Precision-Recall curve

One last useful metric to assess the quality of the model is the **F score**. This measure is simply the weighted average of Precision and Recall. It is a more useful measure than accuracy for our kind of imbalanced data set to assess our test's accuracy.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Since our Precision is low, the F1 score is only 50%, which means that this model would perform unreliable results in the accuracy of its predictions.

In Table 2.1 are the discussed measures given when doing predictions on the training set (where our model was trained on) and the test set:

	Train set	Test set
Accuracy	0.872	0.674
AUC	0.957	0.663
AUPRC	0.915	0.401
Precision	0.722	0.420
Recall	0.920	0.607
F1	0.809	0.497

Table 2.1: Classifier’s scores on train and test set

Instead of doing predictions with our model, we will look at **how** it classified the training set. Looking at how the trees are built, which feature is used at each split and how often can give us a notion of the **feature importance**. We will note 2 types of importance:

Weight The number of times feature X appears in a tree. Figure 2.18 shows that **slope**, **road popularity**, **cloud cover**, **curvature**, **congestion**, **temperature**, **iri**, **time** and **day** are often used to classify data because they provide more meaningful information to separate *slight* from more *serious* accidents.

Gain The average gain obtained by using feature X. Figure 2.17 shows more mixed results on the amount of information we gain by using a certain feature to separate the two classes. However, the speed limit and the context of hitting a tree give a higher information gain and it comes at no surprise since we’ve already noticed these 2 factors when analyzing the proportions of accidents and their severity in Figures 2.10 and 2.12

2.6 Discussion

There are no clear distinctions between the three classes and building a reliable predictive model will require more observations and more information about the context of the accident. However, some factors causing those accidents can clearly be identified. More traffic, on busy road at peak hour or inside cities causes most of the *slight* accidents. Having more serious accidents on clear roads and with nice weather make us wonder if a major factor of fatal accidents is the risky behavior of certain drivers. This is an observation reported by European authorities as well[26]. People tend to speed more

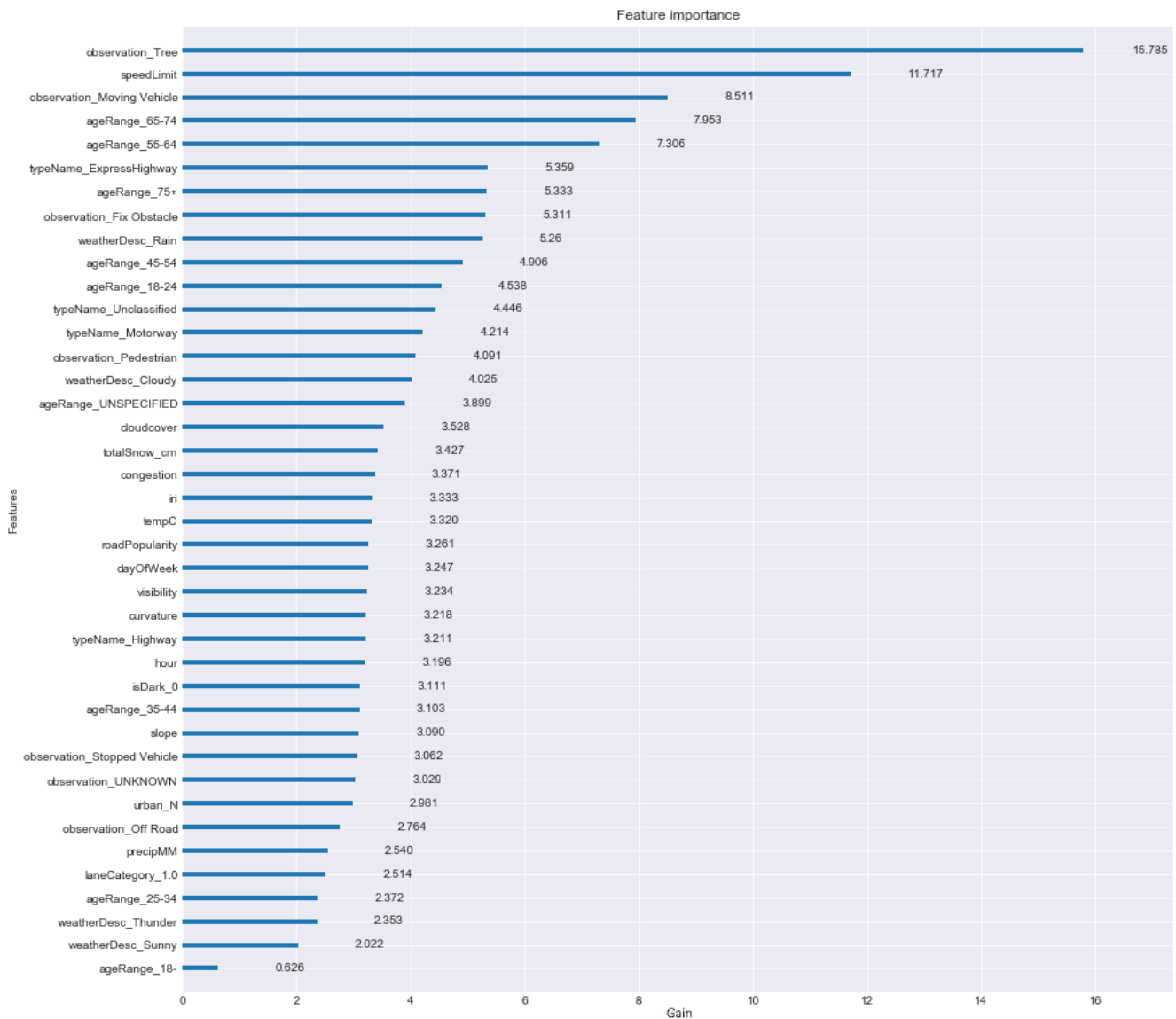


Figure 2.17: Feature Importance based on Information Gain

on road going downhill and as seen in this analysis, slope is an important factor of the seriousness. *Fatal* accidents share a higher ratio on roads marked with *NO OVERTAKING* signs, people not respecting this rule have a higher chance of a frontal collision causing more *serious* accidents.

Such observations come with little to no surprise. Speed radars are often placed on roads where it is tempting to drive fast, i.e. on straight roads and/or going downhill. If there is an increase in the rate of *fatal* accidents where cars should not be overtaking, it is logical to place regulatory signs preventing it.

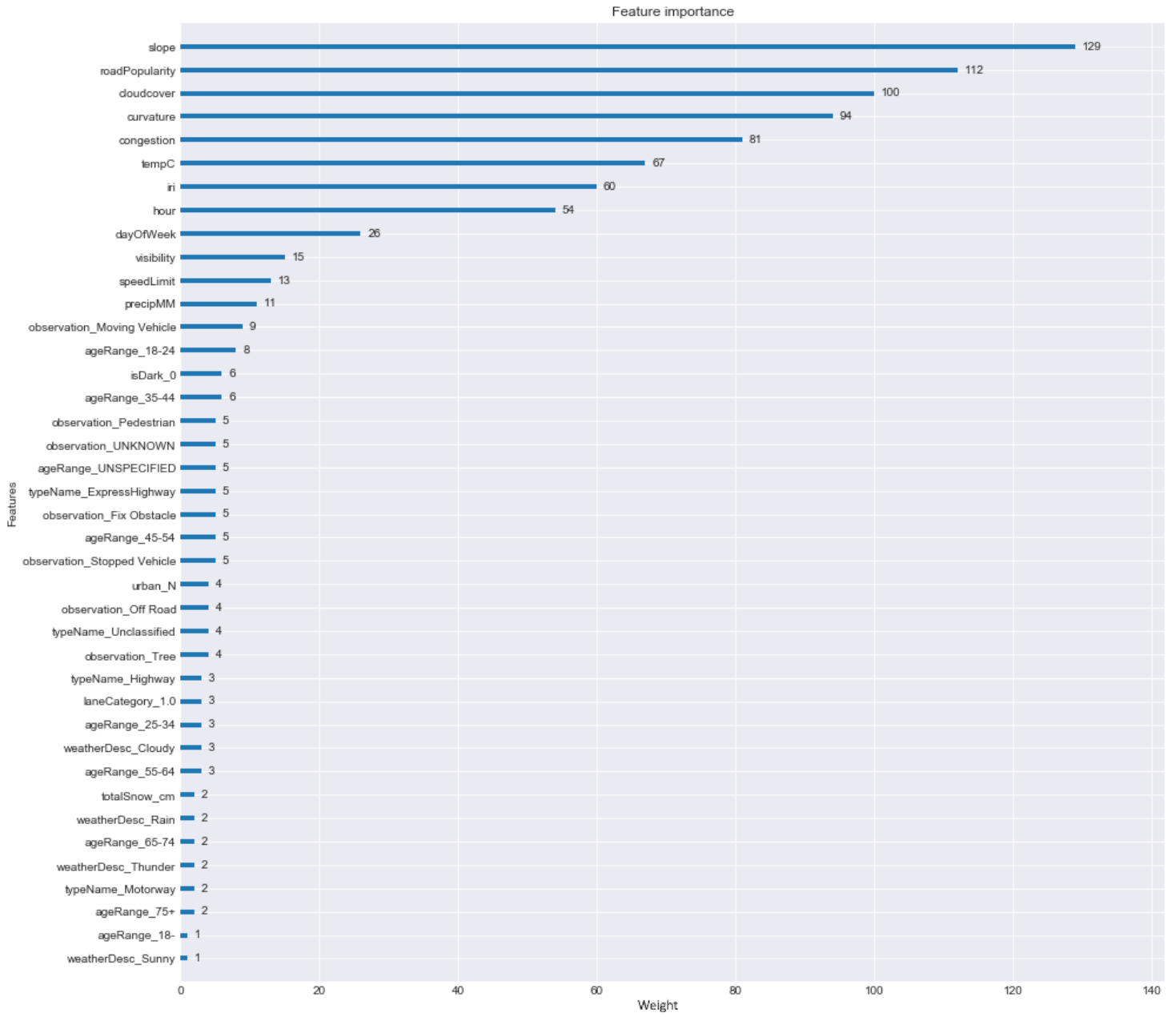


Figure 2.18: Feature Importance based on Weight

Chapter 3

Predictions

As seen with the decision trees built in the previous section, predicting the severity of an accident will be a difficult task. However, there is no reason to limit this task as a classification one. The dataset may well be labelled with 3 types of severity but since the distinction is hard to make, we could think about a model that assigns a probability for each class instead of classifying one. As requested by Motion-S, we should be looking to find metrics that indicates the danger of a road.

What was performed on their side was to find the number of accidents that happened on the trip from A to B. Let's say that 5 *slight*, 1 *serious* and 2 *fatal* accidents were reported on any road composing this trip, the metrics of danger of this trip would be $\frac{5}{8}$, $\frac{1}{8}$, $\frac{2}{8}$. This is not ideal.

First, it indicates that there is a higher "chance" of *fatal* accident than *serious* one which is not true for most if not all roads, and we have previously seen that *fatal* accident are nine times less likely than *serious* ones and this would not be reflected by such measure. Second, if there are no accidents reported on any road composing an itinerary, this doesn't mean that there are no chances of an accident.

What we will be looking at in this chapter, is to find a way to classify our dataset, and in parallel perform a regression to find metrics indicating if road A presents a higher risk of *serious* accidents than road B. Even if they both only have a history of *slight* accidents.

3.1 The models

Similarly to decision trees previously explained, we will use supervised learning. Meaning we will start from the labelled dataset to build a model capable of predicting the output classes. We will then calibrate a confidence level for each severity. For the later, there

exists a method called *Platt's scaling*[27]. It was first introduced in the context of *Support Vector Machines*[28] [**SVM**] but its application has been widened to other classifying models[29].

In this chapter, we will apply the scaling to SVM after having introduced the two notions.

3.1.1 Support Vector Machines

Before the popularisation and improvements of **Neural Networks**, SVMs were the reference for classification models in Machine Learning. They serve for regression and classification tasks but are more widely used for the later.

The model consists in finding the hyperplane that best separates the classes by maximizing the margins between it and the closest observations[30].

- The distance from a point $\mathbf{x}$ to a hyperplane $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + w_0$ is given by $r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$
- The equation of g stays unchanged if we multiply it by a constant $c > 0$. Hence, by fixing the scale, we define the functional margin $|g(\mathbf{x})| = 1$
- The margin at a support vector becomes $r = \frac{1}{\|\mathbf{w}\|}$ and so maximizing the margin is obtained by minimizing $\|\mathbf{x}\|$, or minimizing $\frac{1}{2}\|\mathbf{w}\|$
- We search for parameters $\mathbf{w}$, w_0 of a hyperplane minimizing $\frac{1}{2}\|\mathbf{w}\|$ under the constraints that $\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0 \geq 1$ if $y_i = 1$ and $\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0 \leq -1$ if $y_i = -1$, or simply $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1, \forall i = 1, \dots, n$
- Once the parameters found, the decision function to classify observation $\mathbf{x}$ is

$$f(\mathbf{x}) = \text{sign}[\langle \mathbf{w}, \mathbf{x} \rangle + w_0]$$

This is illustrated in the simple case of separating two classes of a two dimension set in Figure 3.1(a). The solution is rather straight forward if the classes are linearly separable and the model will always provide the optimal separation in this case. Note that we only need to know the support vectors when classifying a new observation, making this model pretty computational efficient once built.

However, in most cases we can't draw a linear decision boundary (b). To solve this, new dimensions can be computed in order to find a hyperplane capable of separating the classes (b). Then we can map this decision boundary back to our original two-dimension

plane (c).

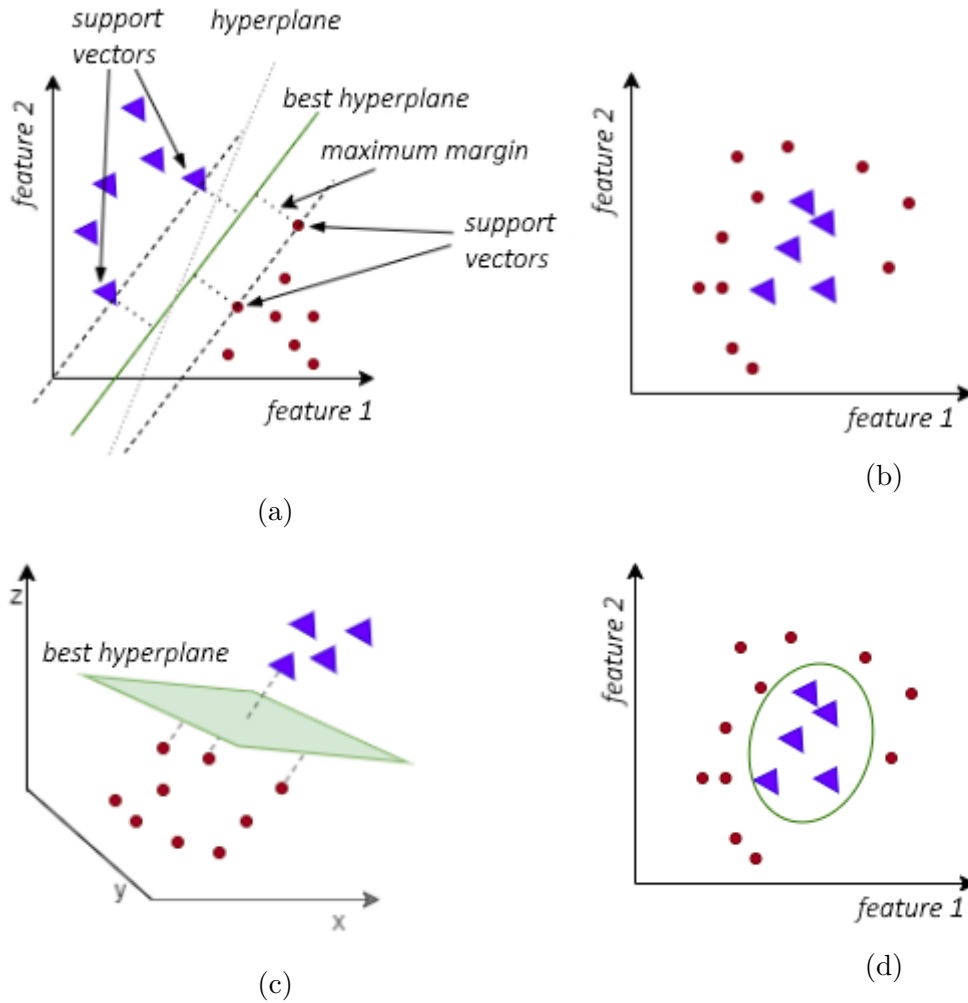


Figure 3.1: Support Vector Machines

This clever way of mapping the original space to a higher one is what makes this model interesting. Calculating new dimensions can be pretty computationally expensive if done for every vector in an already high dimension. But SVM doesn't need the actual vectors, it can get by only with the *dot products* $\langle \mathbf{x}, \mathbf{x}' \rangle$ between them.

This is called the *kernel trick*[31]: Any learning algorithm that uses the data only via dot products can rely on this implicit mapping by replacing $\langle \mathbf{x}, \mathbf{x}' \rangle$ by $k(\mathbf{x}, \mathbf{x}')$

There are many kernels available, most common ones are

Linear $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$

Polynomial $k(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^b$ with $b \in \mathbb{N}, c \geq 0$

Radial Basis Function **[RBF]** $k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right)$ with $\sigma \neq 0$

Sigmoid

$$k(\mathbf{x}, \mathbf{x}') = \tanh(\kappa \langle \mathbf{x}, \mathbf{x}' \rangle + \phi) \text{ with } \kappa > 0 \text{ and } \phi < 0$$

Support Vectors are implemented and available in the python library `scikit-learn` as SVC objects¹. It allows a fine tuning of meta-parameters such as:

Kernel The shape of the decision boundaries as illustrated in Figure 3.2

Degree If we choose a polynomial kernel, specifies its degree

Cost How much a SVM should be allowed to *bend* with the data. A low cost aim for a smooth decision surface and a higher cost aim to classify more points correctly. Note that a higher cost favors over-fitting as shown in Figure 3.3.

Gamma If we choose a RBF kernel, signifies the influence of points either near or far away from the hyperplane. A model with a low gamma will be too constrained and include all points of the training dataset, without really capturing any shape that it could be represented as. A model with a high gamma will better capture the shape of the dataset.

Note that a higher gamma favors over fitting as shown in Figure 3.3.

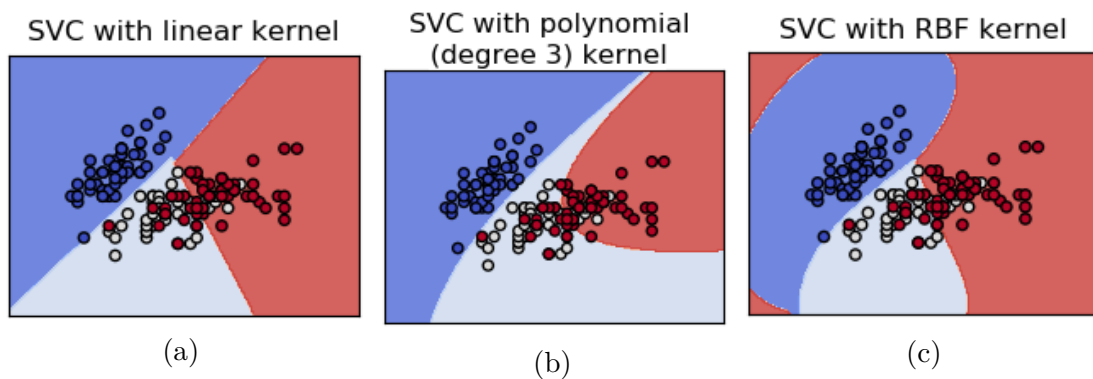


Figure 3.2: Kernels in Support Vector Classifiers @scikit-learn.org

¹<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

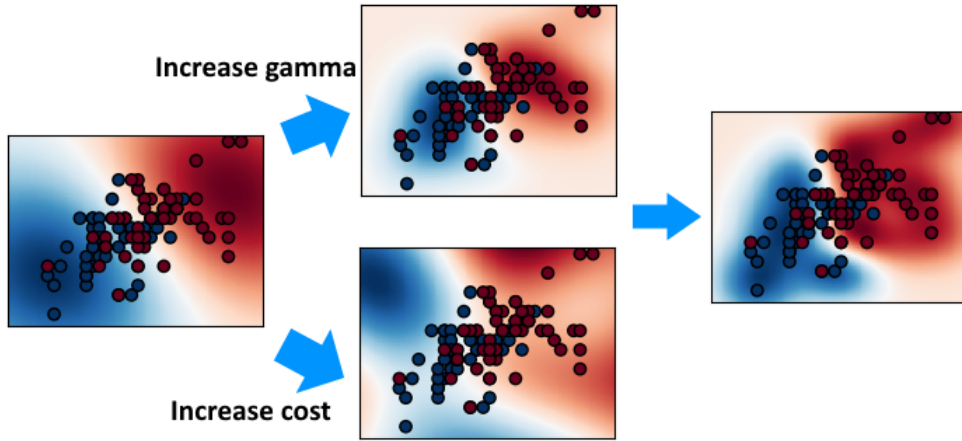


Figure 3.3: Different cost and gamma on a SVM classifier with a RBF kernel @scikit-learn.org

3.1.2 Platt's scaling

Instead of returning a class, we fit a logistic regression to the classifier's scores in order to return the probabilities $P(class|input)$

To do so, we create a new dataset with one dimension being the output of the SVM and the same class label. We then train on this dataset by feeding the output of the SVM as input of a calibration method, which returns a probability. In Platt's case, we are essentially just performing a logistic regression on the output of the SVM with respect to the true class labels.

$$P(y|x) = \frac{1}{1 + \exp(Af(x) + B)}$$

with $f(x)$ as the classifier's score of observation x , and A and B are two scalar parameters learned by the algorithm using a maximum likelihood method that optimizes on the same training set as that for the original classifier f [32].

To illustrate what we want to achieve, let's imagine a dataset with two classes easily separable in Figure 3.4. We start by building the SVM model to retrieve decision boundaries (a), then apply the scaling to provide a nuance of probabilities based on the SVM results (b).

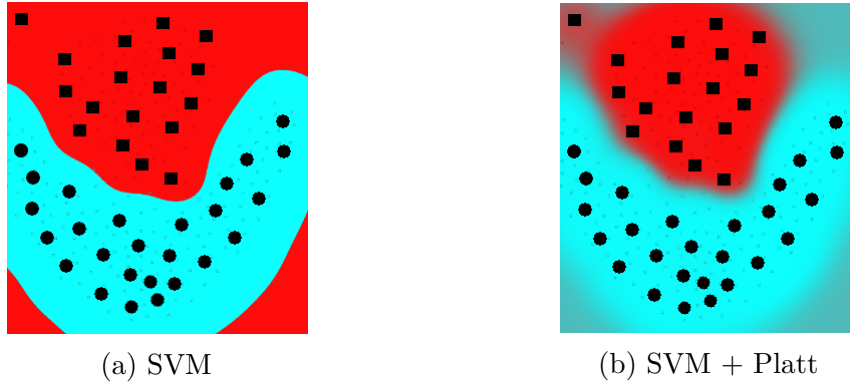


Figure 3.4: Using Platt's scaling on SVM result [33]

3.2 Training & Testing

Before building the model, we divide the dataset into 70% training, 10% validation and 20% testing set.

Now, with the help of the python implementation of SVM given by the library `scikit-learn`, we can start computing the model and tuning its parameters that maximizes classification scores with the validation set.

We consider Radial Basis Function and polynomial kernels, tuning respectively its gamma or its degree, and various different costs.

The tables 3.1 and 3.2 combine the different results provided by the combination of different parameters. Based on accuracy and especially the average f-score of the three classes, we will build our model using a RBF kernel with `gamma=0.1` and `cost=1.5`.

cost	degree	Accuracy (train)	F1 (train)	Accuracy (Validation)	F1 (Validation)
0.1	2	0.82	0.85	0.59	0.35
0.1	3	0.88	0.87	0.6	0.37
0.1	4	0.95	0.92	0.59	0.36
0.1	5	0.99	0.97	0.57	0.35
1	2	0.89	0.88	0.57	0.35
1	3	0.93	0.90	0.59	0.37
1	4	0.97	0.96	0.6	0.36
1	5	0.99	0.99	0.57	0.35
1.5	2	0.89	0.90	0.57	0.35
1.5	3	0.95	0.94	0.6	0.37
1.5	4	0.98	0.98	0.6	0.36
1.5	5	0.99	0.99	0.57	0.35

Table 3.1: Training of SVM using a polynomial kernel with various degree and cost

cost	gamma	Accuracy (train)	F1 (train)	Accuracy (Validation)	F1 (Validation)
0.1	0.01	0.71	0.27	0.70	0.27
0.1	0.1	0.71	0.27	0.70	0.27
0.1	1	0.71	0.27	0.70	0.27
1	0.01	0.71	0.27	0.70	0.27
1	0.1	0.78	0.45	0.71	0.33
1	1	0.85	0.53	0.77	0.42
1.5	0.01	0.71	0.27	0.70	0.27
1.5	0.1	0.82	0.65	0.79	0.52
1.5	1	0.95	0.88	0.75	0.37

Table 3.2: Training of SVM using a radial kernel with various gamma and cost

3.3 Results

As seen on the table 3.2, the model selected doesn't provide a good classification of accidents. This is confirmed by the confusion matrix 3.3 obtained when making predictions on the test set. Due to the class imbalance, the model favors attributing *slight* severity to most accidents. From this table we can obtain the scores in Table 3.4. The average F1-score is 0.55, far from a decent classification model.

		Actual		
		<i>Slight</i>	<i>Serious</i>	<i>Fatal</i>
Predicted	<i>Slight</i>	631	145	23
	<i>Serious</i>	19	92	3
	<i>Fatal</i>	0	1	5

Table 3.3: Confusion Matrix SVM on test set

Class	Precision	Recall	F1-score
<i>slight</i>	0.79	0.97	0.87
<i>serious</i>	0.81	0.39	0.52
<i>fatal</i>	0.83	0.16	0.27

Table 3.4: Classification score with SVM

However, we are looking for the probabilities provided by Platt's scaling and can thus overlook the fact that we are not achieving a perfect classification. Doing so for the test set, we obtain results looking like Table 3.5.

Id	<i>Slight</i>	<i>Serious</i>	<i>Fatal</i>	Predicted	Actual
1	0.754	0.216	0.029	0	0
2	0.758	0.216	0.024	0	0
3	0.583	0.375	0.041	1	0
4	0.752	0.216	0.031	0	0
5	0.743	0.219	0.037	0	1
6	0.458	0.498	0.043	1	1
7	0.486	0.463	0.050	1	1
...	...	...	...	...	...
n-2	0.351	0.389	0.259	2	2
n-1	0.528	0.362	0.108	1	2
n	0.586	0.276	0.137	0	2

Table 3.5: probability distribution over classes per observation

Even when we misclassify an accident, we underline that the road in question has features that increase the probabilities of other severity of accidents. This is illustrated by the results in Table 3.6. When looking at the average probabilities in correctly classified samples, we obtain the first three lines. But looking at all the cases where we made the wrong predictions, for example we have predicted *slight* when it was in fact *fatal*, we see that the average probabilities differ. Comparing the first and fourth line (where the predictions are *slight*), we see that when the model misclassifies, it still gives a higher probability for *serious* and *fatal* on average.

<i>avg P(slight/x) avg P(serious/x) avg P(fatal/x)</i>		
<i>x is 'slight' and the prediction is 'slight'</i>		
0.74	0.23	0.03
<i>x is 'serious' and the prediction is 'serious'</i>		
0.52	0.43	0.05
<i>x is 'fatal' and the prediction is 'fatal'</i>		
0.62	0.28	0.1
<i>x is 'fatal' but the prediction is 'slight'</i>		
0.67	0.26	0.07
<i>x is 'fatal' but the prediction is 'serious'</i>		
0.56	0.38	0.06

Table 3.6: Average probabilities in 'predictions' versus 'actual'

It makes sense to have a higher probability of low severity since, in any situation, a *fatal* accidents could have been *slight* but not always the other way around. Most *serious* and *fatal* accidents have happened on highways at high speed but even on these roads, there were more *slight* accidents. Almost no one died on urban roads limited to 50km/h so certain types of roads don't necessarily allow the possibility of having more *serious* accidents.

With this sort of model, we see that roads with a higher *serious* or *fatal* probability have more chance of having experienced grave accidents.

These are the metrics that *Motion-S* could use to assess the danger of a trip: Trip T is composed of x roads, we feed these roads to the SVM model to retrieve the three probabilities for each road and then average for each class.

To illustrate the concept, imagine having to do a quick trip from the scientific park of Louvain-La-Neuve to decathlonTM. There are two itineraries shown in the map 4.4

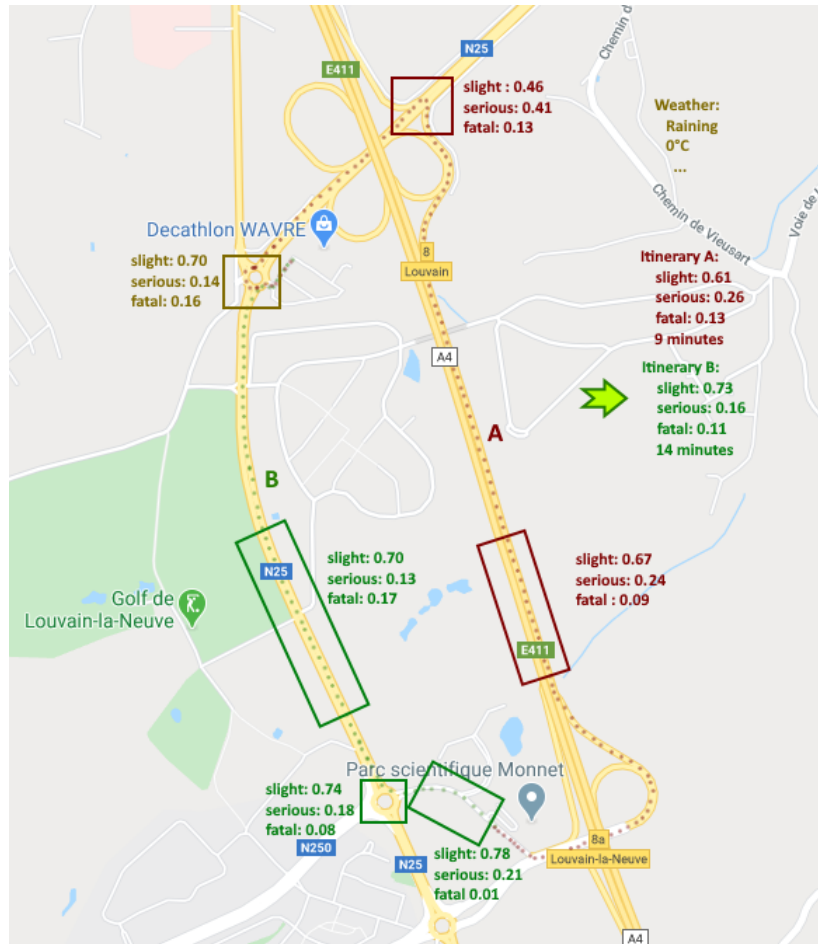


Figure 3.5: Practical use of danger metrics

It is raining and freezing and road B has a lot of traffic at this hour making road A five

minutes faster. The navigator would take all these parameters into account and compute the metrics of danger associated with each itinerary. The driver preferring to be cautious in these conditions, will take the advice that an accident on itinerary B is likely to be less serious than an accident on itinerary A, and thus choose the safer but more time consuming road.

3.4 Improvements

Instead of approximating the severity of an accident, we could try to predict the probability of having an accident to start with. This would make our dataset drastically unbalanced if we consider that all roads that have had accidents have for the major time no accidents. This would require to know the topology of every road in Luxembourg, their occurrence (how many highways? How many 30km/h zones?, etc) and a deep knowledge of the total traffic day to day in the country.

It goes without saying that some of this information would be really hard to get but finding a way to approximate such predictions would be valuable and interesting.

Chapter 4

Data Visualization

In the last chapter of this paper, we will be covering ways to represent and visualize our data. As Machine Learning technologies or Data science in general are getting more and more popular, we frequently see terms such as *Big Data*, *Internet of Things*, etc promising rich data allowing us to better understand the world we live in and to introduce possible ways to make our live easier.

A key requirement to reach any goal is to understand the data that we have collected and find interesting patterns, even when we don't really know what we are looking for.

It is also worth noting that not everyone has a deep understanding of data science or is trained in manipulating data. However this should not prevent them to have access to the information that all of this collected data could provide.

Principles of data visualisation can help with the situations stated above. We will start by formulating a definition and lay out some motivations to use it. Then we will apply a visualization to our accident dataset by extending the functionality of a Visualization tool developed at *LIST*. Finally, we will end this chapter by presenting other visualizations that can be used for different contexts and for different points of view.

4.1 Theory

4.1.1 Definition

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.[34]

4.1.2 Motivations

Here are some reasons that Data Visualization can reveal itself as really useful:

- Exploring** Before any data analysis, it is important to understand the data we have by doing data exploration, as we did in the second Chapter of this paper. A proper visualization of the data can speed-up this exploration.
- Opening Questions** Sometimes, we may want to analyze a dataset without having a clear idea of what to look for. This can be the case when we are looking for patterns in data that we are not aware of, for example. Data Visualization can allow us to see our data from different angles and potentially help us notice some interesting behaviors worth studying.
- Keeping Focus** From the start to the end of a project, visualization can be performed: before the creation of models to provide understanding ; during an algorithm creation to refine, debug, and set parameters ; during deployment of the model to build trust and monitoring.
- Presenting Results** As stated at the beginning of this chapter, not everyone is a data scientist. Presenting data and results to uninitiated readers can be challenging if we only rely on statistical results and presenting the outputs of a model. Instead, we can put these results into context by presenting the whole data under different aspects, compare results visually, and making it easier to grab the attention to important notions.

4.1.3 Techniques

Before visualizing our dataset, we need to understand its type and what kind of data it is composed of. For the latter, we can differentiate five main ones:

Items Attributes Links Position Grids

And separate four common types of dataset seen in Figure 4.1

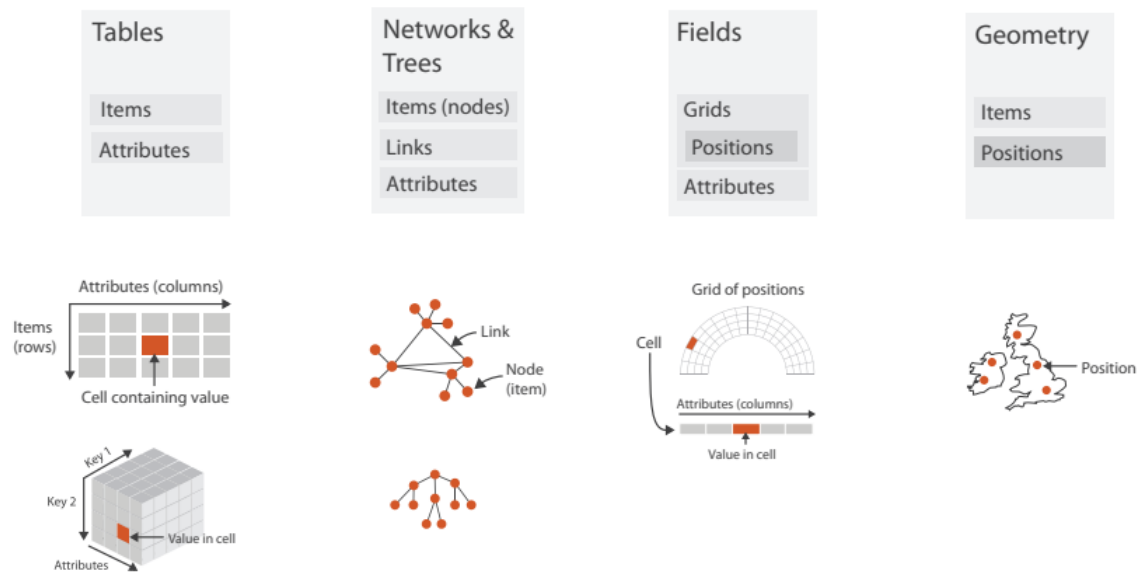


Figure 4.1: Types of dataset and its data @Tamara Munzner

Depending on the type of data that we have, we can choose to visually represent

- Trends, outliers or features
- Topology or paths if it is a network
- Shapes if it is spatial data
- Distributions, correlations, similarities, etc if it is attributes

When we have decided what variables to show and what support to use, we are left with the question of how to encode them.

Continuous features can be represented by points, lines, areas

Categorical features can be drawn as shape or different color or a grading of one.

Numerical features can be represented as different sizes, angles, volumes, or even grading of a color.

Spatial features can be placed on a map or in a network

Combining representations together, with the help of different colors, juxtaposition or superposition can easily allow to compare data and highlight results.

4.2 ICoVeR

ICoVer is a visualization tool¹ developed by a team of researchers at *LIST* in the context of biology to explore genetic data. It helps visualizing binning of metagenomics data, allows comparing results of multiple automated binning techniques and provides means to analyze bins in more detail. Its interactive visualization features help to get deeper insight in the structure of bins and interaction between bins.[35]

4.2.1 Overview

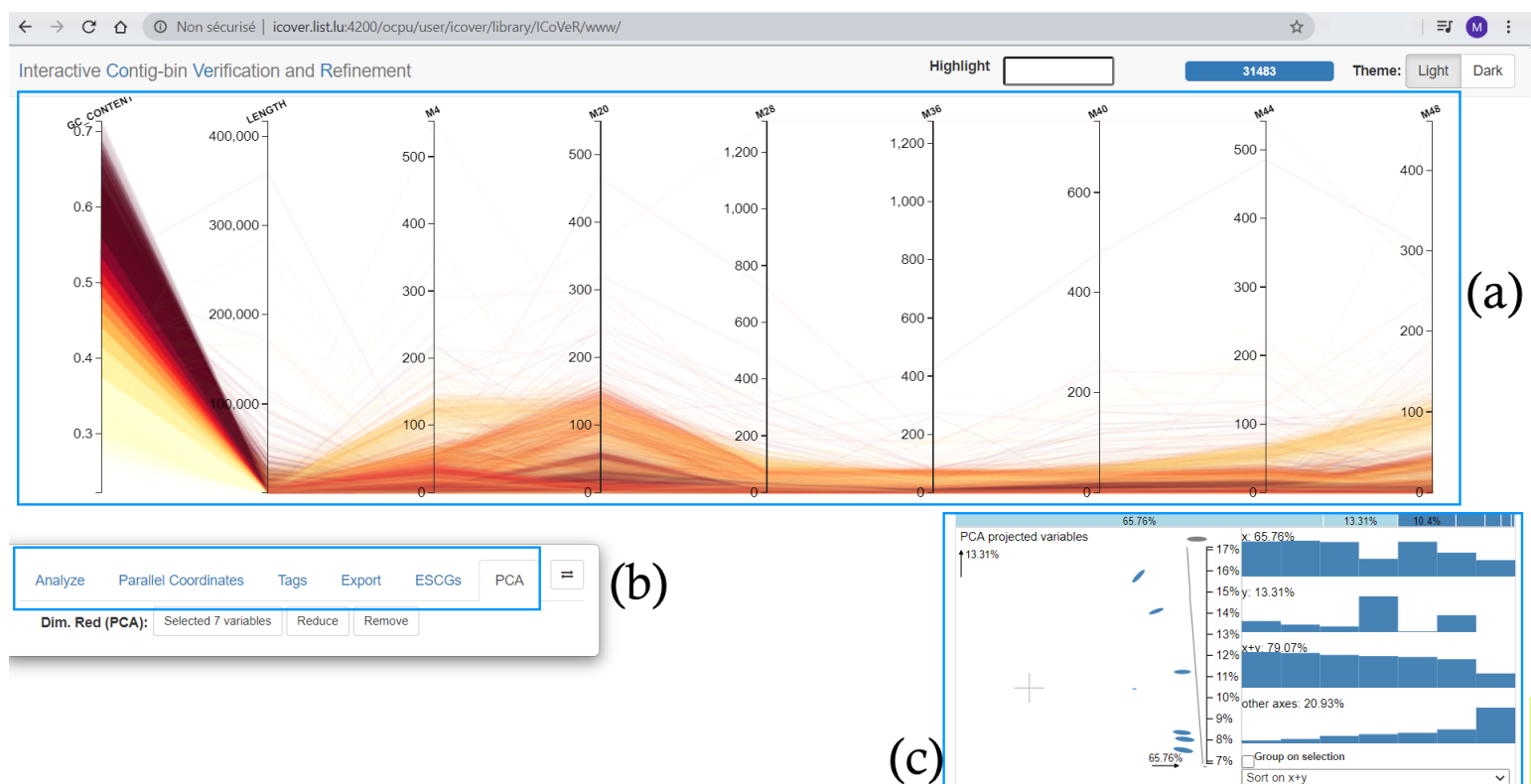


Figure 4.2: ICoVeR at <https://git.list.lu/eScience/ICoVeR>

Its current version is only designed to work with unlabelled dataset having continuous features, mostly found on biological data, see Figure 4.2. The main visualization tool is a barplot showing selected features with gradient of colors depending on a target feature (a). The data can easily be refined and manipulated and transformed by tools such as clustering and dimensionality reduction (PCA)(b). This allow to look at the data at a different angle and refine the dataset accordingly (c).

¹<http://icover.list.lu:4200/ocpu/user/icover/library/ICoVeR/www/>

4.2.2 Improvements

Another aspect of my internship at *LIST* was to allow this tool to accept other types of data such as mine, a labelled dataset, with multivariate features.

The first step was to convert categorical features (`road_type`=`[Highway, Urban, Motorway, ...]` for example) as the tool only works with continuous values. One way to do it is to one-hot encode each concerned variable, i.e. explode each category as binary (for example convert weather forecast to `weather_raining` $\in \{0, 1\}$, `weather_sunny` $\in \{0, 1\}$, etc).

Then implement a transformation technique that can apply to such dataset. Since PCA is already present and that we have seen in a previous chapter that it is similar to LDA, and that LDA is appropriate to the data that we were given, it made sense to implement it in the tool.

As shown in Figure 4.3, the tool now allows the visualisation of our dataset and its reduction by LDA with barplots for its projection on x and y axis. This confirms the result shown in 2.6 that the 3 classes strongly overlap, and that the low severity category is over-represented and tends to overshadow the 2 other severity.

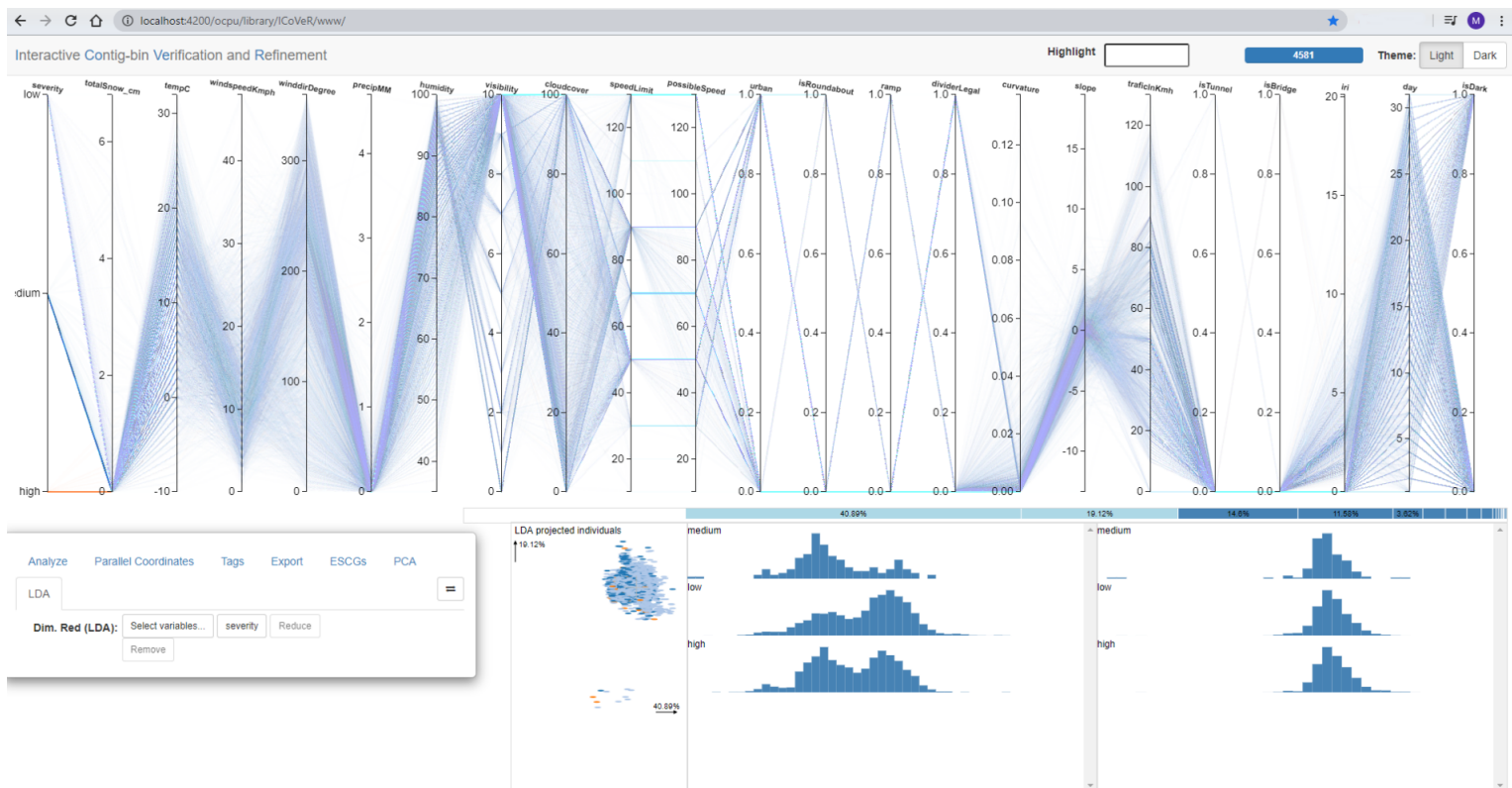


Figure 4.3: Added functionalities for ICoVeR

A user wanting to explore such dataset can now easily select multiple features that could be of interest and have a look at their effects on the target variable or even other features.

A straight-forward example in our case would be that if we select observations that reported at least 1 cm of snow at the time of the accident, the temperature would not be higher than $6^{\circ}C$. Plus we can observe that out of the 270 accidents that happened in these conditions, only 10 were fatal and 60 serious. This means that the proportion of accidents with different severity is still 3%, 22%, and 75% ; same proportion as when all the records are considered. However half of the fatal accidents with snow occurred on urban road with speed limit of 50km/h and the other half on roads with speed limit of 90km/h while in general, less than 30% of *fatal* accident happened on road with a speed limit of 50km/h or lower. We could then make the assumption that urban roads become more dangerous in snowy conditions as opposed to warmer weather.

In addition to the possibility of making custom selections of observations, ICoVeR now includes LDA to visualize the separation of the classes and make it easier to identify outliers. The later is useful to assess if they result of extreme cases or if they are just due to encoding errors.

Performing the linear discriminant reduction with all the features of our accident dataset allows us to quickly single out outliers and analyze them by simply selecting them in the graph and only keep the selection. In our case they concern "exceptional" observations with unusual weather (heavy thunder) with the highest speed limit during daylight thus causing them to stand out.

Unfortunately, since our different classes so strongly overlap, it is hard for patterns in data to be easily noticeable and to bring us more information to what we have already seen at this point.

4.3 Other Visualizations

Data visualization being undeniably useful now in data science and for any decision making in company, professional tools such as *Tableau* started specializing in this field to offer rich and easy ways to represent data.

4.3.1 Maps

The first intuition is to represent each accident on the map of Luxembourg with a color for each severity.

Figure 4.4 gives a clear view on where the accidents occurred in the country. It doesn't take too long to notice that they mostly are regrouped near towns and cities, especially in the capital (a). Looking closer we now see that most accidents happened on major axes rather than on smaller streets (b). Finally, a close-up to individual accidents reveals that in cities, a majority of them take place near or at intersections. An observation that is significantly less represented outside cities

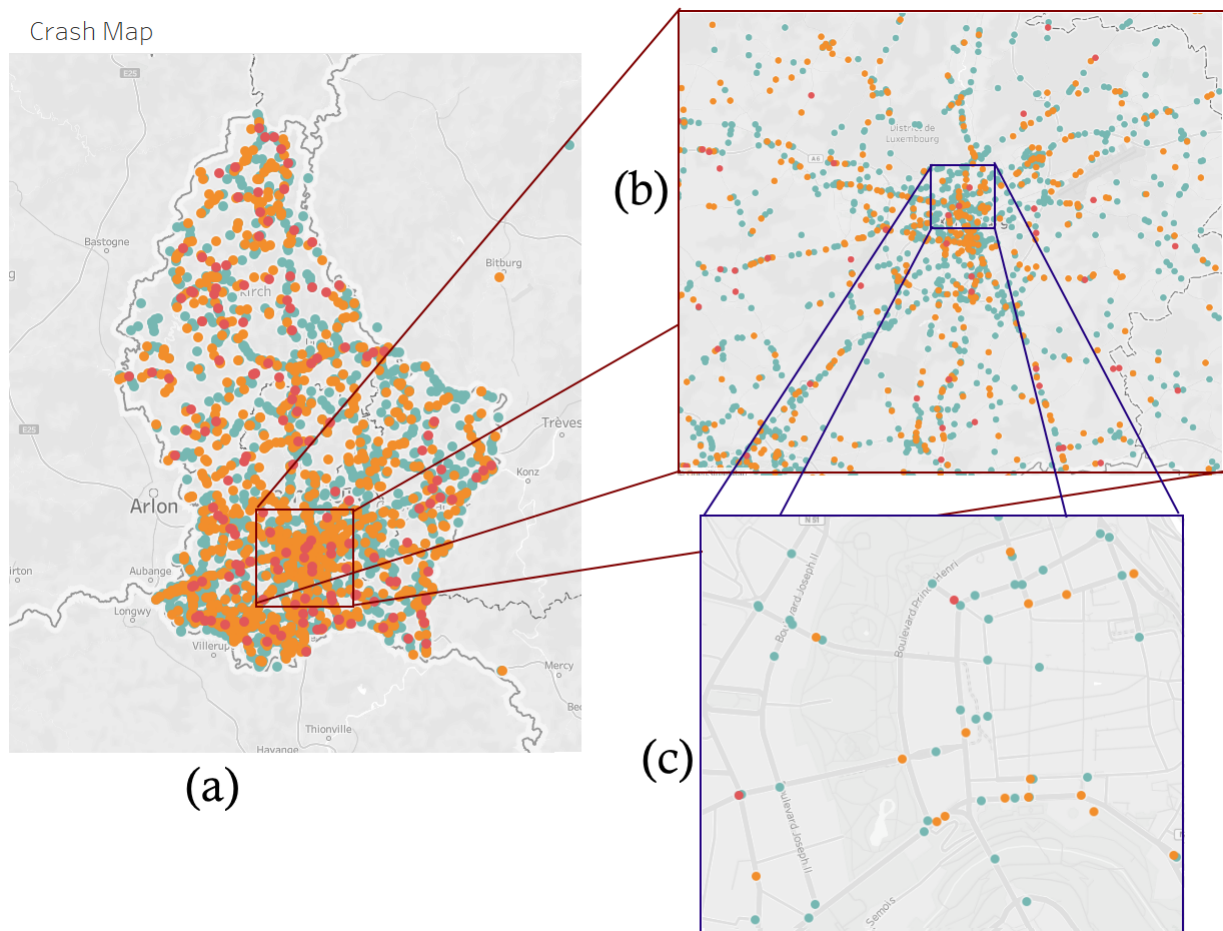


Figure 4.4: Zooms in interactive map of accidents

Another use a geographical maps to visualize our data is to regroup the points seen in Figure 4.4 to their nearest city or region. Using basic principles of visualization, we

can represent the portion of different severity by *pies pieces* in a pie chart and have the size of each pie chart increase with the number of accidents happening within the same city. Figure 4.5 shows such representation and highlights the fact that the vast majority of accidents happened in the south of the country, more populated, but that the ratio of more serious accident increase when looking at less populated regions.

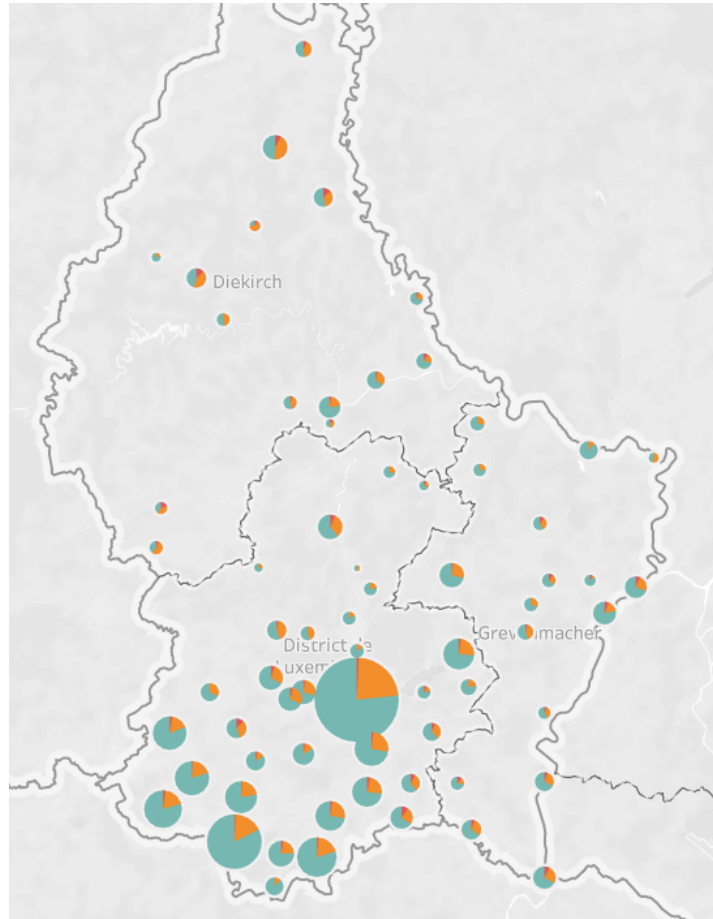


Figure 4.5: Accidents and their severity across cities in Luxembourg.

This technique of mapping information on maps is useful in the context of Geographic Information System[GIS]. GIS[37] is a framework for gathering, managing, and analyzing data. In modern navigation device, we are able to display different layers of information to cover maps. It can be information about the terrain, the vegetation, the country borders, or even the current traffic. Similarly to the traffic layer available with navigation systems such as *Waze* or *Google Maps* that consists of covering segments of road with colors ranging from green or blue to red to inform the current congestion, we could create a new layer informing of the danger of roads. We could use our SVM model defines in the

previous chapter to compute in real-time the current risk present on streets.

Chapter 5

Conclusion

During this research work, we used the power of Machine Learning to gain insights of what affects car accidents. Even though many external factors are involved, and it would be impossible to have all information needed, we still were able to identify danger and calculate danger metrics for roads.

There were three main focuses in this thesis. Firstly, we wanted to **find factors that play an important role** in affecting the severity of an accident. To do so, we modeled the history of accidents with their road and weather characteristics using Decision Trees, generated as a forest with gradient boosting. This technique was expected to bring better insights than a simple statistic observation of individual features by repeatedly trying to correctly classify the data using the most impactful variables. Secondly, we needed a way to **measure the potential severity** of accidents to determine if a certain road is safer than another. We also used a classifier algorithm for this task, using Support Vector Machines, then fitted a logistic regression to return probabilities of classes instead of the classes themselves. Finally, we introduced the basics of Data Visualization to find ways to **graphically represent our data** and how we can benefit from it.

The first model, obtained by Decision Trees, predicted the test set with an accuracy of 67.4%. It had however a low precision of 42%, meaning that it tended to overestimate the severity, while compensating with a recall of 60% resulting in a classifier that correctly predicts six *serious* accidents out of ten. What was interesting for us with this model was to look at how the trees were constructed and what features were the most selected as carrying a better power of separating the classes. We have seen that the speed limit plays an important role in deciding the severity. It is not surprising that a higher speed limit allows more *serious* accidents. Older drivers (54+) are increasingly more prone to have these *serious* accidents, as is hitting a tree. Crashes involving a moving vehicle however

are rarely *serious* or *fatal*. Our measure of road popularity has a high impact as well. Driving with a lot of traffic around often result in *slight* collisions. Slope and curvature are also important factors. A straight road going downhill increase the potential severity as opposed to a curvy road requesting more caution by the driver.

The second model, obtained by Support Vector Machines with a Platt's scaling calibration, allowed us to compute metrics of *danger* for any road. Even if the classifier favored *slight* accidents due to the classes being imbalanced and overlapping, it still was able to identify situations that could cause more serious accident and reflect it with a higher probability for such severity. Given these metrics, one can assess if a road is safer than another by looking at what would be the most likely severity if a crash was to happen there.

Finally, we have seen that Data Visualization is a practical way to have a good overview of our dataset and potentially find patterns or suspicious behaviors. In our case it helped us directly identify outliers. But Data Visualization can be more than that. With proper tools and techniques, it can help people of any background to understand the information that is carried by the data, where the accidents happened, what are the proportion of accidents on different situation, what are the safer roads, etc ...

Throughout this work, we demonstrated that simple yet powerful Machine Learning algorithms can be used to model complex and incomplete data such as traffic accident data. This can be the base for more complete models, with more observations and context, to deeply analyze mobility data and improve the safety on our roads. And with the growing interest in self-driving cars; understanding the difficulty of driving, and identifying the dangers on roads are a key to successful future ways of transportation.

Improvements and Future Works

We could explore the very first dataset provided by *Motion-S* regrouping hundreds of thousands of trips, and compute the *danger* metrics given by our SVM model for all roads encountered and analyze how drivers behave depending on them. Are they more cautious where roads are marked with a higher probability of severe accidents? Or, on the contrary, are they driving more carelessly, explaining the higher degree of *serious* accidents? Our work mainly focused on spatial risks but it would be interesting to correlate it with behavioral risk.

Additional information could be searched to enrich the data. Finding if an accident happened near an intersection would certainly be of great value, but it turned out to be

difficult to obtain in this research. Other factors could be added as well, for example eventual distractions on the road or police controls nearby, or even the position of the sun potentially blinding drivers.

Many other Machine Learning techniques could also be considered to analyze risks on roads. Some use clustering algorithms to group similar roads and then look at what severity is more represented per cluster[38], other apply pattern mining algorithms to find combinations of road characteristics affecting a change in severity[39]. A full study comparing the performances of different techniques could provide insights on what concept is the most suitable and for what goal.

Finally, we could expand the use of this data to other purposes. Instead of the severity, we can predict the potential cost of accidents happening on certain roads, useful for insurance companies. Or we can find ways to estimate the risk of an accident happening, helping for a better dispatch in medical interventions, better road signalization, etc.

Personal Achievements

I learned a lot writing this thesis. It first helped me to gain my first professional experience by allowing me to do an internship on the topic. Then it gave me the occasion to face a real-world problem with the tools that I had acquired throughout my academic years. It also made me realize that results are sometimes hard and frustrating to find, and that reality can be far more complex than 'simple' situations that are used to explain Machine Learning concept. But curiosity, imagination, and scientific rigor allowed me to to successfully carry out this work.

This showed however the importance to understand the data we are analyzing and not simply applying concepts as a *black box*. This proves the great value that data scientists can bring to the research community and society itself.

I do believe in the many possibilities that Machine Learning, Artificial Intelligence, Big Data, IoT, etc. could bring to the world. There is information everywhere; using it and understanding its meaning is the key to improve and facilitate our life.

Bibliography

- [1] *Analyse des marches automobiles* CORAM, Avril 2014
http://www.coram-research.com/wp-content/uploads/2014/06/8_Vehicles-in-use-2014-ppx-resultats_fr.pdf
- [2] *Security and Safety of transport in the EU*, European Commission 2019,
https://ec.europa.eu/transport/themes/security_en
- [3] *Road safety Measures in the EU*
https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1793
- [4] *Towards Zero, Ambitious Road Safety Targets and the Safe System Approach*
International Transport Forum (2008) OECD
- [5] *"A data mining approach to characterize road accident locations"*, Sachin Kumar, Durga Toshniwal
J. Mod. Transport. (2016);24(1):62–72
- [6] *Predicting injury severity levels in traffic crashes: a modeling comparison*, Abdel-aty, M.A., Abdelwahab, H.T., 2004.
J. Transp. Eng. 130, 204–210.
<http://refhub.elsevier.com/S0001-457530286-5/sbref0010>
- [7] *Analysis of traffic accident severity using Decision Rules via Decision Trees*, Abellan, J., Lopez, G., Ona, J.De, 2013.
Expert Syst. Appl.: Int. J. 40, 6047–6054.
<http://dx.doi.org/10.1016/j.eswa.2013.05.027>
- [8] *Using support vector machine models for crash injury severity analysis*, Li, Z., Liu, P., Wang, W., Xu, C., 2012.
Accid. Anal. Prev. 45, 478–486.
<http://dx.doi.org/10.1016/j.aap.2011.08.016>

- [9] *Rapport d'Accidents corporels de circulation Routière*
Plate-forme de Données Publique Luxembourgeoise
<https://data.public.lu/fr/datasets/accidents-accidents-corporels-de-circulation-routiere/>
- [10] *Here Technologies*
<https://www.here.com/>
- [11] *Types of Data Sets in Data Science, Data Mining Machine Learning*, Tarun Gupta ,
Towards Data Science 2019
<https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a>
- [12] *Road Attributes Available, Here.com*
https://tcs.ext.here.com/pde/attributes?region=WEU&release=20105&url_root=pde.api.here.com
- [13] *Calculus: An Intuitive and Physical Approach*(2nd ed.). p. 458
Kline, Morris
- [14] *Little Book of Profiling*, Sayers, M.W.; Karamihas, S.M. (1998)
University of Michigan Transportation Research Institute
- [15] *Encoding cyclical continuous features - 24-hour time*, Ian London (2016)
<https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/>
- [16] *PCA versus LDA* Martinez, A. M.; Kak, A. C. (2001)
IEEE Transactions on Pattern Analysis and Machine Intelligence. 23: 228–233
doi:10.1109/34.908974
- [17] *Discriminant function analysis* Garson, G. D. (2008)
NC State Univeristy
<https://chass.ncsu.edu/garson/pa765/discrim.htm>
- [18] *Data mining with decision trees: theory and applications*, Rokach, Lior; Maimon, O. (2008)
World Scientific Pub Co Inc. ISBN 978-9812771711
- [19] *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*, Stacey Ronaghan
Towards Data Science, 2018

<https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>

- [20] *Top-down induction of decision trees classifiers-a survey*, Rokach, L.; Maimon, O. (2005)
IEEE Transactions on Systems
doi:10.1109/TSMCC.2004.843247.
- [21] *Bagging Predictors*, Breiman, Leo (1994)
University of California Berkeley
<https://www.stat.berkeley.edu/~breiman/bagging.pdf>
- [22] *Boosting with AdaBoost algorithm*, Emer, Eric; MIT
<http://math.mit.edu/~rothvoss/18.304.3PM/Presentations/1-Eric-Boosting304FinalRpdf.pdf>
- [23] *On the momentum term in gradient descent learning algorithms*, Qian, Ning (1999)
doi:10.1016/S0893-6080(98)00116-6
- [24] *An Introduction to ROC Analysis*, Fawcett, Tom (2006)
Pattern Recognition Letters. 27 (8): 861–874
doi:10.1016/j.patrec.2005.10.010
- [25] *Prevalence Threshold and the Geometry of Screening Curves*, Balayla, Jacques (2019)
doi: <https://arxiv.org/abs/2006.00398>
- [26] *Speed choice: why do drivers exceed the speed limit?* European Commission (2020)
Mobility and Transport: Road Safety
https://ec.europa.eu/transport/road_safety/specialist/knowledge/speed/many_drivers_exceed_the_speed_limit/speed_choice_why_do_drivers_exceed_the_speed_limit_en
- [27] *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers* Platt, J. C. (1999). (pp. 61–74). MIT Press.
- [28] *Using SVM weight-based methods to identify causally relevant and non-causally relevant variables*, Statnikov, Alexander; Hardin, Douglas; Aliferis, Constantin; (2006)
Vanderbilt University

- [29] *Predicting good probabilities with supervised learning*, Niculescu-Mizil, Alexandru; Caruana, Rich (2005)
ICML. doi:10.1145/1102351.1102430
- [30] *Machine Learning: Classification and Evaluation*, Pierre Dupont
UCLouvain
- [31] *Pattern Recognition*, Theodoridis, Sergios (2008)
Elsevier B.V. p. 203. ISBN 9780080949123.
- [32] *Choosing multiple parameters for support vector machines*, Olivier Chapelle; Vladimir Vapnik; Olivier Bousquet; Sayan Mukherjee (2002)
doi:10.1023/a:1012450327387.
- [33] *Probability Calibration* Machine Learning with JSAT, 2013
<http://jsatml.blogspot.com/2013/06/probability-calibration.html>
- [34] Industrial visualization tool *Tableau*™, 2020
<https://www.tableau.com/learn/articles/data-visualization>
- [35] "ICoVeR-an interactive visualization tool for verification and refinement of metagenomic bins"
Bertjan Broeksema, Magdalena Calusinska, Fintan McGee, Klaas Winter, Francesco Bongiovanni, Xavier Goux, Paul Wilmes, Philippe Delfosse, Mohammad Ghoniem.
BMC Bioinformatics 18, no. 1 (2017): 233.(10):891–921, 1905.
- [36] *Visualization Analysis Design*, Tamara Munzner, University of British Columbia, Feb 2020
<https://www.cs.ubc.ca/~tmm/talks.html>
- [37] *Geographic information system: Old principles with new capabilities*, Maliene V, Grigonis V, Palevičius V, Griffiths S (2011)
Urban Design International
doi:10.1057/udi.2010.25
- [38] *Kernel density estimation and K-means clustering to profile road accident hotspots*, Tessa K. Anderson
Accident Analysis and Prevention 41 (2009) 359–364
- [39] *A data mining framework to analyze road accident data*, Kumar and Toshniwal
Journal of Big Data (2015) 2:26
DOI 10.1186/s40537-015-0035-y

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl