CAR ACCIDENTS IN SWITZERLAND: A REVIEW OF LEADING FACTORS

IBM Data Science Professional Certificate – Capstone Project

August 2020

Cédric M. G.

Table of Contents

1. Introduction

1.1 Background

While a reduction in number has been observed over the last decades, car accidents are still counted in thousands in Switzerland in 2020. Because the direct and indirect consequences of such events (injuries, death, psychological damages, material damages, etc.) are sizeable, there is value in identifying what are the causes of the accidents so that adequate prevention measures can be put in place. Moreover, it would be valuable to society - not the least from a resource planning standpoint - to understand when accidents are most likely to occur, and respectively what outcome severity (light injuries, severe injuries, fatal outcome) can be expected depending on when and under what circumstances the accident took place. Since 1992, the Swiss Federal Statistics Office (OFS) is collecting data on car accidents country-wide and making such information available to the public. This analysis will leverage this data.

1.2 Problem

The objective is to explore a year 2019 dataset from the Swiss Federal Statistics Office (OFS) and determine what are the key factors that drive the outcome of an accident for the involved car(s)' passengers: light injuries, severe injuries, fatal outcome. Additionally, the outcomes of this analysis can be used as a prescriptive tool to :

(1) Have the appropriate medical emergency resources allocated for the times, locations and circumstances when accidents are most likely to occur, with a particular emphasis on the severe and life-threatening cases.

(2) Design prevention measures based on those accident factors identified as having the largest influence on accident outcomes.

1.3 Interest

By being able to allocate medical emergency resources more efficiently and by being able to reduce injuries and deaths through prevention campaigns, society as a whole will reduce the economic impact of road hazards. This analysis is therefore aimed at decision-makers of the Swiss Confederation, notably those in charge of Transportation and Medical Affairs. Beyond economic considerations, there is also a moral value in reducing the suffering and deaths of the thousands of people affected by road accidents.

2. Data Sources and Cleaning

2.1 Data Sources

The dataset used here is defined as "Road accidents where at least one of the parties was injured or worse". As a result, this dataset does not report on material damages or other consequences than bodily injuries. It is also worth noting that this dataset does not distinguish between what exact type of vehicle was involved, whether a car, bicycle, motorcycle, tractor, pedestrian, skater, etc.

The dataset used in this analysis was obtained from the website of the Swiss Federal Statistics Office (OFS). A data browser allows the user to select the dimensions and time range of interest, within the limits of the Office's data structure. The used time range is the calendar year 2019. No other data source was used.

The variables in this dataset are as follows. A detailed description follows after the list:

Variables in the dataset

Types of accidents : TYPE_ACCIDENT

Type of road: TYPE_ROAD

Severity of the accident: SEVERITY (the dependent variable)

Month of the accident: MONTH

Day of the week: DAY

Time of the accident: TIME_ACCIDENT

Types of accidents

SKID: The vehicle went into a skid/sideslid and/or the driver lost control of the car.

OVERTAKE : While trying to overtake or changing lanes. This also includes the variable state where the accident happened when the vehicle was returning to its original lane.

TURNING: While the vehicle was turning to change directions, ie. enter a new road.

INTERSECTION: Accident taking place at a crossroad or junction of two roads with the two implied vehicles staying on their respective roads.

BACK: The vehicle crashed into the back of another vehicle that was either mobile or immobile.

PARKING: While getting in or out of a parking spot.

ANIMAL: Accident created by an animal.

FRONTAL: Frontal collisions.

PEDESTRIANS: Accidents involving one or several pedestrians.

OTHERACC: Other types not captured in the categories above.

Time of the accident

NIGHT : Between midnight and 6am

MORNING: Between 6am and noon

AFTNOON: Between noon and 6pm

EVENING: Between 6pm and midnight

Day of the week

The days were grouped into two categories : Weekday (WEEKDAY) and Weekend (WEEKEND)

Month

For simplicity, the data has been grouped into seasons which are representative of average road conditions.

WINTER: Dec-Feb Icy roads

SPRING: Mar-May Can be misty or rainy

SUMMER: Jun-Aug Generally good road conditions

FALL: Sep-Nov Slippery fallen leaves and rains

Severity of the accident

LIGHT_INJURIES: Light injuries to at least one of the involved parties.

SEVERE_INJURIES: Severe injuries to at least one of the involved parties.

DEATH: Death of at least one of the involved parties.

Type of road

The following road types have been grouped in two buckets representative of their respective maximum speed.

HWY: Highways, semi-highways and similars. The speed limit is typically 120 km/h, respectively 100 km/h for the semis.

NONHWY: Main roads, secondary roads and other roads. The speed limit is typically 80 km/h or less.

2.2 Data Cleaning

The data was extractable only in a form where each row represents a unique combination of explanatory variables (such as time, day of the week, type of road, etc.). There are then 10 columns corresponding to the years 2010-2019, where it is reported in each column how many accidents happened in the said year for the given set of explanatory variables of the row.

Data cleaning consisted first in the following basic steps :

(1) Translating the labels from French to English. This was performed in Excel directly on the CSV file.

(2) Shortening the labels, for easier coding purposes. This was performed in Excel directly on the CSV file.

(3) Transforming the non-numerical classifiers into dummy variables usable by the Machine Learning model discussed below. This was performed in the Jupyter Notebook.

(4) Because a time-evolution is not the primary concern of this analysis, it was decided to focus the analysis on the year 2019 (as opposed to using the full dataset tracing back to 1992), as it is most likely more representative of the current road conditions and car technology. This dataset contains over 17'000 accidents, hence is a fairly significative dataset. This was performed in the Jupyter Notebook.

(5) Itemize into individual rows the rows where the number of occurences in superior to 1. (so that each row represents an individual accident).

(6) Group certain variable values into more general bins, as described in section 2.1. This was performed in Excel directly on the CSV file.

2.3 Feature Selection

For this analysis, all features in the set were preserved, with the note that some values were grouped as explained just above.

3. Methodology

3.1 Exploratory Data Analysis

The exploratory data analysis was conducted by analysing the leading causes of accidents.

This was done using all the classifiers to be able to identify which are more prone to creating accidents.

The results are presented in the Results section.

3.2 Machine Learning Approach

This is a typical classification project, where instances have to be classified in 3 possible classes (death, severe injuries, light injuries) using certain classifiers.

Because the classifiers and the target at hand are non-numerical, it was decided to use a decision tree. The independent variables had to be transformed into floats (dummy variables approach).

KNN was not selected because it requires numerical independent variables to make sense.

Similarly, regressions and logistic models were not picked, as the first one requires continuous values (or at least a discretization) and the second requires a binary target at least in its simplest forms (ie. without softmax methodology).

3.3 Code for the ExploratoryAnalysis and ML Approach

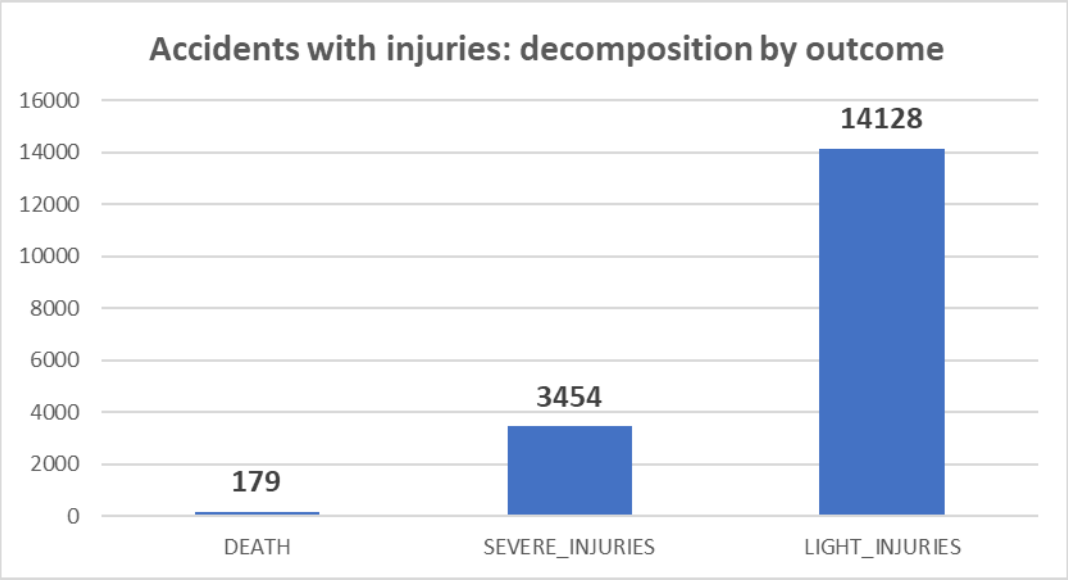
The code is presented on the Jupyter notebook that has been shared at the address

https://github.com/CGIBM/Coursera_Capstone/blob/master/Coursera_Capstone.ipynb

4. Results

4.1 Traditional Analysis methods

One starts by decomposing the total number of injury-bearing accidents as follows. One can see that deaths are a minority of the outcomes and that light injuries are the most common outcome. There is a total of 17'761 accidents in the dataset for 2019.



We then move on to decompose the accidents by their cause. Skids (loss of control) are by far the leading accident factor, suggesting that inattention, poor judgement of road conditions or understanding of vehicle physics are all areas of potential improvement.

Back collisions are the next leading cause of accidents, one that can be attributed most probably to inattention and sudden congestions at high speeds, such as on the highway.

Finally, overtaking is the 3rd cause of accident, suggesting that people overestimate the physics of their vehicle or the sufficiency of a given visibility.

Accident Causes	Death	Severe injuries	Light injuries	Grand Total
ANIMAL	1	27	41	**69**
BACK	7	284	3'281	**3'572**
FRONTAL	14	159	471	**644**
INTERSECTION	10	139	691	**840**
OTHERACC	1	39	115	**155**
OVERTAKE	13	474	2'516	**3'003**
PARKING	3	38	213	**254**
PEDESTRIANS	35	551	1'546	**2'132**
SKID	90	1'499	4'208	**5'797**
TURNING	5	244	1'046	**1'295**
Grand Total	**179**	**3'454**	**14'128**	**17'761**

Regarding the days of occurrence of accidents, the following table demonstrates that accidents occur proportionately more in the weekdays than during the weekends. This might be linked to work-home commuting.

Accident day, scaled on a per-day basis	Death	Severe injuries	Light injuries
WEEKDAY	0.5	42.2	9.6
WEEKEND	0.2	12.1	3.7

As demonstrated in the table below which shows the occurrence decomposed by quarters of the day, the above hypothesis seems to be validated as there are more accidents occurring during the afternoon (defined as noon until 6pm) than during the other periods of the day. The prevalence of the afternoon is particularly strong with almost 45% of all accidents taking place during that time. Work fatigue could be an explanation. One can see that the morning is the second most important time, hence further strengthening the hypothesis of work-home commuting. This does not mean that the work-home commuting is more dangerous than other trips, but this is when most of the traffic occurs.

Accident time	Death	Severe injuries	Light injuries	Grand Total
AFTNOON	77	1'528	6'352	7'957
EVENING	38	748	2'847	3'633
MORNING	45	935	4'120	5'100
NIGHT	19	243	809	1'071
TOTAL	**179**	**3'454**	**14'128**	**17'761**

Further breaking down the set between weekdays and weekends, one notices that the weekend has a heavier representation of accidents in the afternoons and nights. One could therefore think that it would be valuable to conduct prevention campaigns about the underlying phenomenons occurring at these times of the day (alcohol and fatigue during the nights and leisure commuting during the weekend[1]

Looking at the seasons, one could initially think that many accidents occur in winter in Switzerland, due to the hazardous road conditions. However and as the table below demonstrates, the summer is actually the most dangerous period.

Accident by season	Death	Severe injuries	Light injuries	Grand Total
FALL	51	844	3'796	4'691
SPRING	35	810	3'273	4'118
SUMMER	56	1'262	4'239	5'557
WINTER	37	538	2'820	3'395
Grand Total	**179**	**3'454**	**14'128**	**17'761**

It is also interesting to see that the weekend is particularly heavy in terms of skid-generated accidents. One would need to understand whether this is driven by alcohol (over-confidence, poor judgement) or other factors.

Type of accident vs. day of the week	Week day	Week-end	Grand Total
ANIMAL	0.4%	0.2%	0.4%
BACK	21.4%	15.9%	20.1%
FRONTAL	3.6%	3.6%	3.6%
INTERSECTION	5.1%	3.6%	4.7%
OTHERACC	0.9%	0.8%	0.9%
OVERTAKE	17.8%	14.1%	16.9%
PARKING	1.5%	1.2%	1.4%
PEDESTRIANS	13.2%	8.2%	12.0%
SKID	28.5%	46.0%	32.6%
TURNING	7.6%	6.4%	7.3%
Grand Total	100.0%	100.0%	100.0%

Looking at the causes of accidents during those same seasons, one sees that skids are uniformly the leading factor. Surprisingly, they are not overly represented in the winter but are actually more prevalent in the summer, casting a question on what behaviors lead to such skids.

[1] For instance, there is a very large number of cars (and a lot of congestion) on the highways in summer and winter on the Sundays when people come back from the mountains and other leisure destinations in the country.

Row Labels	ANIMAL	BACK	FRONTAL	INTER-SECTION	OTHER ACC	OVER-TAKE	PARKING	PEDE-STRIANS	SKID	TURNING	Grand Total
FALL	0.4%	21.2%	3.2%	4.7%	0.9%	17.8%	1.3%	12.9%	29.9%	7.7%	100.0%
SPRING	0.3%	21.9%	3.8%	4.6%	0.9%	17.9%	1.7%	11.6%	30.4%	6.9%	100.0%
SUMMER	0.5%	17.0%	4.0%	4.8%	0.8%	16.7%	1.6%	8.2%	38.2%	8.2%	100.0%
WINTER	0.3%	21.6%	3.3%	4.8%	0.9%	14.9%	1.0%	17.5%	30.0%	5.7%	100.0%
Grand Tot	**0.4%**	**20.1%**	**3.6%**	**4.7%**	**0.9%**	**16.9%**	**1.4%**	**12.0%**	**32.6%**	**7.3%**	**100.0%**

It is also interesting to see that back collisions and overtaking accidents are less prevalent in the summer (as a % of the total), maybe suggesting that this is due to the visibility conditions which are better during warm and dry days.

Looking at the prevalence of skid accidents, it seems obvious that measures need to be put in place to address such a phenomenon. More precisely, it would be commendable to conduct additional studies to determine what are the sub-factors behind that situation, whether inattention, alcohol, over-confidence or road conditions for instance.

Looking at accidents by type of road, one sees that many accidents still occur outside of highways. With a dense network of highways, one could have expected that Switzerland would have displayed more casualties on these roads than what the numbers below suggest.

Accident by road type	Death	Severe injuries	Light injuries	Grand Total
HWY	20	178	1'632	1'830
NONHWY	159	3'276	12'496	15'931
Grand Total	**179**	**3'454**	**14'128**	**17'761**

4.2 Machine Learning modeling

Here we used a Decision Tree approach with dummied classifiers as all of them were non-numerical.

At first sight, the tree gives a good 79% accuracy. However, looking closer at the data, one realizes that its performance is quite poor. The model is effectively only able to classify the data into "Light injury" buckets and the non-accuracy ends up corresponding to the ratio of severe injuries & deaths to the total of accidents.

After trying with a Gini maximization, the results were exactly the same, at 79%. Changing the tree depth to 2, 4, 5 did not help either. From that standpoint, one can conclude that this ML model is not a success.

One can see particularly that dummying variables such as the accident type (skid, turning, takeover, etc.) does not work as these are strictly discrete categories. And from a conceptual standpoint, it appears normal that one cannot predict with a lot of precision the outcome of an accident with the classifiers at hand.

One would therefore recommend discarding this ML analysis and focusing on the traditional results presented above.

5. Discussion of the Results

At the end of this analysis, it appears clearly that not much can be done with simple Machine Learning techniques with the dataset as provided by the Swiss authorities. Indeed, the categorical classifiers complexify the analysis a lot and one would need more advanced techniques (not taught in this course) to be able to complete a more meaningful analysis.

Conversely, one can also notice that quite a few insights are already obtained from the traditional data analysis methods, and that there is limited interest in exploring a dataset of that type with an ML approach.

6. Conclusion

In conclusion, the reader will have taken note that key factors. of accidents in Switzerland over the year 2019 have been as follows :

- The summer is a critical period.
- Accidents seem to overly take place in the late afternoon and evenings, possibly when people come back from work.
- Skidding is still the number one factor of accidents, suggesting that people get distracted and/or that their judgement of road conditions and vehicle physics should be improved.
- In the weekends, there is an over-proportionate number of skidding-generated accidents, suggesting that measures need to be taken in prevention or those, f.ex. in the areas of alcohol consumption and judgement alteration.

Measures should therefore be implemented in priority against these factors. Additionnally, it would be commendable to organize the allocation of emergency medical resources and personal for those times identified as critical, such as for instance the summer, the weekdays and the afternoons.