Prediction of car accident severity

# Carlo Peano

# September 30, 2020

## 1. Introduction

### 1.1 Background

According to the WHO[[1]](#footnote-1), even though vehicles have become much safer in the last decades, every year around 1.35 million people still die because of a road traffic crash and between 20 and 50 million more people suffer non-fatal injuries with many incurring a disability.

If we take a different perspective and consider the economic impact at national level, road traffic accidents also cost around 3% of gross domestic product to most countries[[2]](#footnote-2).

Therefore, there is a great interest in different parts of society (such governments, decision-makers, carmakers, drivers, insurance companies) in changing and decreasing this trend.

### 1.2 Problem

A solution that would reduce the number of incidents could be the chance to warn a driver about the possibility of getting into a car accident and how severe that incident would be, given the weather and road conditions. In this way people would drive more carefully or even stay home.

Transforming this solution into a machine learning problem, I used a dataset provided by a city and its police department (in our case Seattle City and the SPD - Seattle Police Department) to predict the probability and severity of an accident based on the conditions of weather, light and the road.

## 2. Data

### 2.1 Data understanding

In this project I used the data provided by the SPD (Seattle Police Department) and recorded by the Traffic Records. This dataset - called Data-Collisions.csv - includes all types of collisions involving cars, bikes, pedestrians and others (around 200,000) from 2004 to present.

After I extracted the dataset, I looked at the columns, their meaning and their link to the objective of the research study, i.e. to predict the probability and severity of an accident based on the conditions of weather, light and the road.

Thanks also to the description of the attributes (available together with the dataset[[3]](#footnote-3)), I have been able to define the attributes and the target variable. For me it was obvious to choose SEVERITYCODE (i.e. the severity of the accident) as the dependent variable. SEVERITYCODE is a categorical variable and follows a code that corresponds to the severity of the collision: 2 (injury) and 1 (property damage).

Out of the 37 attributes available in Seattle accident dataset, I chose 8 of them as independent variables, thanks – as previously said – to their logical connection to the objective of our research study.

|  |  |
| --- | --- |
| Variable | Description |
| JUNCTIONTYPE | Category of junction at which collision took place |
| WEATHER | A description of the weather conditions during the time of the collision |
| ROADCOND | The condition of the road during the collision |
| LIGHTCOND | The light conditions during the collision |
| SPEEDING | Whether or not speeding was a factor in the collision |
| LOCATION | Description of the general location of the collision |
| PERSONCOUNT | The total number of people involved in the collision |
| VEHCOUNT | The number of vehicles involved in the collision. This is entered by the state |

*Table 1. Variables and their description*

JUNCTION, WEATHER, ROADCOND, and LIGHTCOND are the main attributes since they are directly connected to the project’s objective.

PERSONCOUNT and VEHCOUNT makes us understand how big the accident can be: an accident can involve a lot of cars and people and still have nobody injured or no property damage.

SPEEDING has always been considered to have a direct impact on the probability of the collision and is the only attribute that is actually a choice of the driver.

Lastly, even though LOCATION may be way too local, and – for this reason – a weak variable in a general research study, it is useful to show the different level of danger in the districts of Seattle city.

### 2.2 Data Preparation

**2.2.1 Data cleaning**

Once I chose the attributes and the target variable, I dropped the unnecessary columns and analysed deeper the necessary ones.

At this point, there were several problems with the dataset.

Firstly, most the eight attributes had missing data because the SPD did not write the data:

1. "SPEEDING" has 185340 missing data
2. "JUNCTIONTYPE" has 6329 missing data
3. "WEATHER" has 5081 missing data
4. "ROADCOND" has 5012 missing data
5. "LIGHTCOND" has 5170 missing data
6. "PERSONCOUNT" has 0 missing data
7. "VEHCOUNT" has 0 missing data
8. "LOCATION" has 2677 missing data

If most of the missing data were easily solvable by deleting them, those ones of SPEEDING seemed immediately problematic, as more than 90% of the data was missing. However, a direct observation of the attribute values showed that the police took in consideration this attribute – by writing “Y” – only when speed was one of the causes of the accident, otherwise they left the variable empty. As a consequence of this, I considered all the missing data as an “N”, i.e. speed was not one of the reasons of the incident.

Secondly, four attributes had data internally classified as “Other” and/or “Unknown” which were a sort of hidden missing data as they did not actually provide a real information about the attribute and could actually provide confusion:

1. "JUNCTIONTYPE" has 9 "Unknown"
2. "WEATHER" has 832 "Other" and 15091 "Unknown"
3. "ROADCOND" has 132 "Other" and 11012 "Unknown"
4. "LIGHTCOND" has 235 "Other" and 13473 "Unknown"

I used to different methods for managing the two parameters. “Other” means that the data cannot be replaced by any other available observation, therefore it was replaced with NaN and its rows were dropped together with the missing data. “Unknown” was replaced with the mode of the attribute so as to avoid biasing the dataset.

Thirdly, there was a problem with PERSONCOUNT (the total number of people involved in the collision) and VEHCOUNT (the number of vehicles involved in the collision. This is entered by the state): PERSONCOUNT has 5544 incidents involving nobody (zero people) and VEHCOUNT has 5085 accidents involving zero vehicles. These observations were not interesting for the project as its objective implies the involvement of people and vehicles, the lack of one or both of them could bias the dataset. Therefore, I drop those rows were PERSONCOUNT and VEHCOUNT were zero.

**2.2.2 Correct data format**

Considering the data format of the chosen attribute (see Table 2. Data format), I did not need to change any of them.

|  |  |
| --- | --- |
| Attribute | Type |
| SEVERITYCODE | int64 |
| LOCATION | object |
| PERSONCOUNT | int64 |
| VEHCOUNT | int64 |
| JUNCTIONTYPE | object |
| WEATHER | object |
| ROADCOND | object |
| LIGHTCOND | object |
| SPEEDING | int64 |

*Table 2. Data format*

**2.2.3 Feature selection**

In order to balance the dataset, I used one hot encoding technique to convert categorical variables to binary variables and append them to the feature Data Frame. Then I defined the feature set X and the labels y and Normalise the data.

## 3. Exploratory Data Analysis

Scatter plot:

PERSONCOUNT

VEHCOUNT

JUNCTION, WEATHER, ROADCOND, and LIGHTCOND

1. "Road traffic injuries", World Health Organisation (WHO), 07/02/2020, <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries> [↑](#footnote-ref-1)
2. Ibid. [↑](#footnote-ref-2)
3. "ArcGIS Metadata Form", <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf> [↑](#footnote-ref-3)