## Description and discussion of the problem

**Road traffic collisions are a leading cause of death in the United States for people aged 1–54[[1]](#footnote-1). In order reduce this problem road planers, law enforcement officers, and others need decision support from data scientists. Vast amount of data exists about the nature of accidents and how they occur, but modern data science techniques most be applied to go from information overload to insight and understanding.**

**This project aims to build a model that predicts, using selected features, what the severity of an accident will be. This model can then be used by responsible authorities to simulate different type of accidents by adjusting features such as weather conditions, road conditions, and if the driver is speeding or not. From these simulations decision makers can select the optimal strategy to**

**As an extra bonus the findings will also be valuable for car drivers, enabling them to predict in which situation they are most likely to be in a severe accident.**

## Description of the data and how it will be used to solve the problem.

In this submission I will be using the Seattle dataset of all collisions from 2004 until present date. The dataset contains 194.673 accidents described with 38 features (before adjusting for potential duplicates). These features include a categorical severity code, geolocation, information on the type of collision, how many where involved in the accident, in what type of junction it occurred, if the driver was under the influence or inattentive, weather, road, and light conditions and more.

The severity code is a categorical variable with 1 representing “property damage”, and 2 representing “injury collision”. This will be the variable that the model will attempt to predict using selected features.

### 2.1 Feature selection

For this model I will be using three features describing internal driver factors:

|  |  |
| --- | --- |
| **Feature name** | **Feature description** |
| INATTENTIONIND | Whether or not collision was due to inattention. (Y/N) |
| UNDERINFL | Whether or not a driver involved was under the influence of drugs or alcohol. |
| SPEEDING | Whether or not speeding was a factor in the collision. (Y/N) |

I will also be using three features that are external to the driver:

|  |  |
| --- | --- |
| 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. |

### 2.2 Data cleaning

As with all real-world data the quality of the data is varying and needs cleaning before applying machine learning algorithms. The first problem to address is that of zero values in the external features (WEATHER, ROADCOND, LIGHTCOND).

Zero values in external features is represented as “Other”, “Unknown”, or empty cells:

|  |
| --- |
|  |
| Figure 1: Showing how many rows have missing values |

Dropping the zero values will reduce the dataset with 24.716 rows, a reduction of 12.7%. It will probably have some effect on the how well the machine learning algorithms perform, but for now I am willing to make the sacrifice. If on a later stage I assess that the performance is to low, I will come back and adjust by randomly assigning values by frequency, or simply choosing the most common value.

The labels in the external features need to be encoded according to how much it effects the driver. For example with feature “WEATHER” I will give “clear”, “partly cloudy”, and “overcast” the value “0”, “overcast” as “1”, “fog/smog/smoke”, “blowing sand/dirt”, and “Raining” as “2”, and “Snowing” and “Sleet/Hail/Freezing Rain” as “3”. Similar conversions are done with feature “ROADCOND” These conversions are based on assumptions and are open to adjustments if the model does not perform optimal.

For internal features (“INATTENTIONIND”, “UNDERINFL», «SPEEDING») the data is cleaned by converting «Y» to 1, «N» to 0. For empty values I will assume that these are negative and will assign «0». Lastly names of labels are changed and will now be addressed as «INATTENTION», «DUI», and «SPEEDING».

### 2.3 Caveat

CRISP-D is a model often represented with feedback loops between phases. Please remember that this is the first attempt at business understanding, data understanding, and data preparation and cleaning. I expect that modelling and evaluation will reveal a need for additional cleaning and/or understanding.

The data is still biased towards accidents being less sever. This problem, and normalisation, will be addressed in the methodology section.

1. https://www.cdc.gov/injury/features/global-road-safety/index.html [↑](#footnote-ref-1)