Car accident severity

Mohamed ZOUIDINE

I. Introduction

Car accident, or car crash, also called traffic collision, occurs when a car come into collision with another car, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Car accident often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.

According to World Health Organisation, every year approximately 1.35 million people die as a result of a road traffic crash. Between 20 and 50 million, more people suffer non-fatal injuries, with many incurring a disability because of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.

A number of factors contribute to the risk of collisions, including vehicle design, speed of operation, road design, road environment, driving skills, impairment due to alcohol or drugs, and behaviour, notably distracted driving, speeding and street racing. These factors can be used to train machine learning models to predict accident severity, and these models will alert drivers to the severity of the situation, which can help to reduce losses,

II. Data

This project serve to build machine learning models to predict the severity of accident car. A dataset appropriate for this project can be found here. The dataset describes the severity of the collision and the factors involved in it for accidents occurring between 2004 and 2018 in the Seattle city.

The dataset contain 221738 samples and 40 features. The target variable is "**SEVERITYCODE**", which is a code that corresponds to the severity of the collision:

- 3 fatality
- 2b serious injury
- 2 injury
- 1 prop damage
- 0 unknown

After we examine the meaning of each feature, it was clear that there was many features describes the collision. In this project, we choose to work just with the features that caused the collision, which are:

- **INATTENTIONIND**: Whether or not collision was due to inattention.
- **UNDERINFL**: Whether or not a driver involved was under the influence of drugs or alcohol.
- **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.

Therefore, we remove all other features. We have now a new dataset with 221738 samples, 6 independent variables (**INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND** and **SPEEDING**) and one dependent variable (**SEVERITYCODE**). In this form, there are several problems with the dataset:

1. **Missing values**: the figure below shows the number of missing values for each feature.

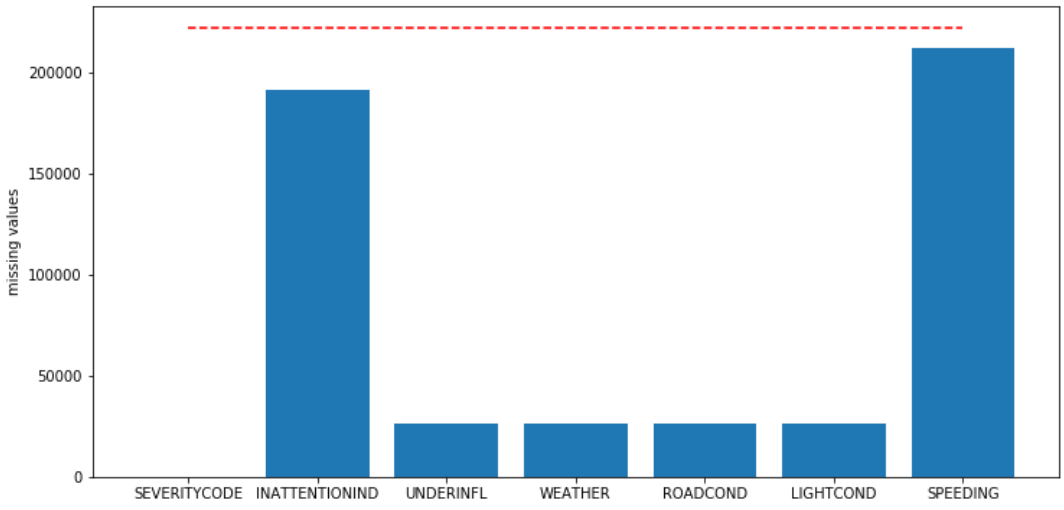

Figure 1: Number of missing values.

As you can see in the figure, the independent variables **INATTENTIONIND** and **SPEEDING** present a lot of missing variables (around 80% and 90% respectively). The solution is to remove these two features. For the other features, we use the imputing method to replace them with the frequent value.