Accident Severity Rating (ASR)

Maris Sekar

September 26, 2020

# **1 Introduction**

## 1.1 Background

Road safety is a major problem that is often taken lightly. The World Health Organization (WHO) reports that “Approximately 1.35 million people die each year as a result of road traffic crashes”. Many people incur a disability as a result of their injury. Many more lives change either directly or indirectly due to road accidents. The personal injury, emotional and mental health impacts are one side of these unfortunate incidents. But there are economic impacts to the individual, their family and their community as well depending on the nature of the accidents. These impacts are often unforeseen and the relationships associated between these impacts are often forgotten as an aftermath to an unfortunate outcome. It is important to be able to assess the accident severity and the accident severity level to incorporate the human fatality, traffic delay, property damage, or any other type of accident parameters that are involved so that the right level of support can be provided to the affected individuals. For example, this can be a standardized and independent rating – Accident Severity Rating (ASR) that can be shared with insurance providers for them to process insurance claims using a logical and structured approach to assess the claim. This can be used in mandatory driver training programs to focus and share awareness of parameters that contribute to improving road safety (for example, we can determine how much more likely someone is to get into an accident if they are speeding) and further reduce road accidents.

## 1.2 Problem

Data such as location, collision type, number of people involved in the accident, number of vehicles involved, incident date, road conditions, weather and whether speeding was a factor may provide great insights when determining the severity level. The objective of this project is to predict the severity level of accidents and determine the major contributing factors to this severity level.

## 1.3 Interest

This can be of interest to the road transport safety communities within the government to better understand the contributing factors and the relationships between them in order to introduce targeted awareness campaigns and programs to reduce road safety incidents. This can be of great interest to insurance companies so that they are able to logically assess a claim based on the logically derived severity rating.

# **2 Data Acquisition & Cleaning**

## 2.1 Data Sources

I used the provided data set named Data-Collisions.csv shows data provided for this project from the Seattle Department of Transportation for the period between 1st January 2004 and 20th May 2020. There was a total of 194,673 collision incidents with 37 attributes. However, there were a lot of missing columns (FATALITIES, INJURIES and SERIOUSINJURIES) when compared to the provided metadata as well as additional columns (REPORTNO and STATUS) that were not present in the metadata. The missing columns are very important in determining the severity. Another major deficiency with this data was SEVERITYCODE column only had two values out of a total of five possible values according to metadata. This made me more uncomfortable to choose this data so I tried to find an alternate dataset.

Fortunately, Kaggle had a better version of this data set provided also by the Seattle Department of Transportation. There was a total of 221,144 collision incidents with 40 attributes. This had more consistent columns including FATALITIES, INJURIES and SERIOUSINJURIES. This data set also had all possible values for SEVERITYCODE which is important because this is the target variable to predict.

## 2.2 Data Wrangling

The data was analyzed and cleansed based on the requirements for this project. I checked for duplicate rows. There are no duplicate rows.

Some fields need to be cleaned up so that Y, N become 0s and 1s – one hot encoding. For example INATTENTIONIND, UNDERINFL and JUNCTIONTYP fields were cleaned up this way.

I removed the following columns, reasons are given on the same line below:

* + EXCEPTRSNCODE & EXCEPTRSNDESC appears to describe exceptions and there are 5638 lines with not enough information or insufficient location information so remove this column
  + PEDROWNOTGRNT – Most are “N” so we are not losing much information by removing this.
  + SDOTCOLNUM – an arbitrary number that is assigned by SDOT.
  + SEGLANEKEY AND CROSSWALKKEY– Mostly has 0. So, we will remove these columns
  + HITPARKEDCAR – Most have “No”.
  + X, Y – don’t need location information for our models.
  + LOCATION (string) – don’t need location information.
  + INCDATE (INCDTTM has date and time) – Don’t need date info.
  + INCDTTM - Don’t need date info. Good for analysis though. Like to look at what time and days most incidents happen.
  + SDOT\_COLCODE – We are already using ST\_COLCODE which seems to contain similar information.
  + SDOT\_COLDESC (Same reason as above)
  + ST\_COLDESC – don’t need description for features.

## 2.3 Feature Selection

The following columns were chosen as features:

* ADDRTYPE (string)
* INTKEY
* COLLISIONTYPE
* PERSONCOUNT
* PEDCOUNT
* VEHCOUNT
* PEDCYLCOUNT (Not included because pedestrian/person similar)
* JUNCTIONTYPE
* WEATHER
* ROADCOND
* LIGHTCOND
* SPEEDING
* INATTENTIONIND
* UNDERINFL
* ST\_COLCODE

Target is the SEVERITYCODE field