Road Vehicle Accident Severity Prediction in Seattle, WA

Gabriel Siqueira Kakizaki

October 5, 2020

1 Introduction

1.1 Problem and Background

Road vehicle accidents are a problem that in 2019 caused more than 38 thousand estimated deaths, and injuries in about 4.4 million people, only in the USA. In this project we will explore the use of machine learning models to predict the accident severity, using open data provided by the city of Seattle.

1.2 Stakeholder Interest

Policymakers need information on what factors cause road accidents, especially the severe ones, when creating or improving on existing preventive policies. Machine learning models can help extract the needed insights from the data.

2 Data

2.1 Data sources

The city of Seattle makes data about road vehicle crashes available, under the Public Domain Dedication and License (PDDL). The most recent dataset can be downloaded from here.

2.2 Data description

Downloaded data had information about approximately 195 thousand road vehicle collisions. Examples of some relevant features are latitude and longitude coordinates, date and time of occurrence, weather, road and light conditions, among others.

To prepare the data for analysis and modeling, first we removed the entries missing the location, which represented 2.74% of the total rows. Second, we removed columns which missing values represented more than 10% of the total values, although some of them could have had a good predictive value (if the driver was speeding, for example). Third, we imputed the remaining missing values with the most common value for each of the columns.

Another problem that we had, common in real world classification projects is the balance of the target variable. In our binary classification problem, approximately 70% of the cases represented a low severity collision (property damage), while only 30% were a high severity collision (with injury). This will be solved in the modeling pipeline.