Predicting car accident severity

1. Introduction

1.1 Background
Road accidents result in both injuries, material damage and in the worst-case loss of life. These accidents could in many cases be prevented if more caution was exercised. By understanding what the factors are that contribute to road accidents and their severity many accidents may be avoided through preventive measures.

1.2 Problem
Despite efforts to make roads and cars safer the number of accidents is still too high. Legislators, transportation agencies and car manufacturers all try to make traffic safer for all. However, in order to effectively prevent road accidents there is a need to understand the factors contributing to accidents. By utilizing data science to better understand what factors and their importance in contributing to accidents we can help reduce the number of accidents in the future. There will be common factors that correlate and contribute to road accidents more than others. By identifying these a model for preventing future accidents can be developed.

1.3 Interest
Legislators, transportation agencies and car manufacturers will be interested in understanding what factors that contribute to accidents and determine their severity in order to make their preventive work more efficient. It could also in the interest of drivers to understand what factors increase the likelihood of accidents in order to improve their driving and take precautionary measures when faced with accident prone situations.

2. Data acquisition and cleaning

2.1 Data source
I used the shared dataset available Clicking here. The dataset includes all types of collisions, factors which can be included in the machine learning model and the accident severity. The dataset contains many observations and various attributes (38 columns and 194673 rows), which makes it good for solving our problem.

2.2 Data cleaning

There were some missing data where the accident severity was unknown labelled as severity 0, these values were removed to gain better accuracy.

2.3 Feature Selection

There were many features in the dataset that are not relevant to determining the accident accident severity. However, the following features were selected for the machine learning algorithm due to their likely effect on accident severity.

ADDRTYPE - Collision address type:
-Alley
-Block

-Intersection

ROADCOND – the road condition when the accident occurred.

LIGHTCOND – the light condition when the accident occurred.

WEATHER – the weather condition when the accident occurred.

INATTENTIONIND – Was the driver not paying attention (yes or no).

UNDERINFL - Was the driver under the influence (yes or no).

SPEEDING - Was the driver speeding (yes or no).