Predicting the severity of car accidents with Machine Learning

Data Section

The dataset we use for this project was downloaded from the coursera web page. The initial .csv file contains 194,673 data points about car collisions in the city of Seattle beginning in the year 2004. Attributes covered in the dataset are the conditions oft he weather and the road, as well as the collision type and the number of fatalities.

Data Cleaning:

The original dataset contains 37 attributes with a lot of missing values. So to work with it we need to clean the data up. Doing so we dropped all rows with missing values, leaving us with 110,586 collision records.

Feature Selection:

To efficiently train our model we need to focus on just a couple of features. In this case we decided to concentrate on: SEVERITYCODE, JUNCTIONTYPE, PERSONCOUNT, WEATHER, ROADCOND, LIGHTCOND. These features seemed to best most suitable for training a model. The target feature will be the SEVERITYCODE.

Feature Engineering and Data Balancing:

To efficiently work with the selected features we have converted all categorical values to numerical values, by using the df.replace() method.

Furthermore we sampled down the majority class of the target feature to avoid a bias in the model. To do so we used the resample method.