Predictive Analytics of Car Accidents in Seattle

2. Data

2.1. Description

The data set used in the present study is provided through the Coursera Capstone Course link. The raw data set contains car accident information in Seattle from 2004 to 2020. The raw data set contains 194,673 entries of accidents with 38 descriptive features.

2.2. How the Data Will be Used

The data set has some duplicate entries. We first drop the duplicated rows. This reduced the data set to 189,542 rows.

Some of the entries also contain Nan values. For example, two of the features that we consider as independent variables, road condition and weather condition, contain 5012 and 5081 Nan values, respectively. We will remove the Nan values for the exploratory data analysis and predictive modeling. In the exploratory data analysis phase, we will mostly be focused on analyzing pairs of selected features. Thus, in this phase, we will drop Nan values in the subsets that consist of the selected features for the exploratory analysis. In the later predictive modeling phase where we apply a machine learning approach, we will drop all Nan values in the dataset that is used by the machine learning algorithm.

A further examination of the data shows that some features contain redundant information or information that is not very useful for the purpose of the present analysis. Below we list the columns we drop from the raw data set and brief descriptions of the reasons why they are dropped. All other features that are not listed below are kept.

INCKEY: dropped because we use the OBJECTID as the unique identifier

COLDETKEY: dropped because we use the OBJECTID as the unique identifier

LOCATION: dropped because location data can be obtained through X, Y coordinates

EXCEPTRSNCODE: exception codes that are not needed for the analysis

EXCEPTRSNDESC: descriptions of the exception codes that are also not needed

SEVERUTTDESC: dropped because information is contained in SEVERITYCODE

HITPARKEDCAR: dropped because COLLISIONTYPE reflects whether a parked car is involved

SDOT_COLDESC: dropped because the information is available in SDOT_COLCODE

ST_COLDESC: dropped because the information is available in ST_COLCODE

CROSSWALKKEY: dropped because there is uncertain meaning regarding this key number