Prediction of Accidents' Severity Based on Collisions

Introduction/Business Problem section

The automobile has become an important part of daily life. It influences how we live, work, and how we spend our time. It helps you to go to the workplace on time, sends your kids to school, and takes you and your family to a vacation on a beach. However, at the same time that it brings happiness, it also brings danger – car accidents. According to the statistics, there are 16,438 car crashes per day in the U.S. alone [1]. On a global scale, there is approximately 3287 death every day due to car accidents [1]. Generally, people would try not to drive at rush hour, bad weather, or dark place to avoid the chance to get into an accident. However, sometimes driving in a bad condition is unavoidable. For example, your kids are waiting for you to pick them back home after the piano class on a snowy night. This raises the question that how we maintain our daily needs and avoid accidents and related injuries at the same time. Thus, it is necessary to provide a model to predict the severity of an accident, thus one could estimate the dangers of different driving behaviors and stay safe.

Herein, we would build a machine learning model based on collision data to predict the accident severity. Data is provided by SPD and recorded by Traffic Records downloaded from the IBM course on Coursera. The provided dataset will be cleaned, processed, and used to build a classification model to predict whether an accident will lead to property damage or injury.

Data section

The data has total 37 attributes and 194673 instances. According to the meta data form, we are going to first remove the data without any explanation. Thus, columns, OBJECTID, INCKEY, COLDETKEY, REPORTNO, EXCEPTRSNCODE, EXCEPTRSNDESC, STATUS, SEVERITYCODE.1. We only need to keep SDOT_COLDESC or SDOT_COLDCODE, ST_COLDESC or ST_COLDCODE, SEVERITYDESC and SEVERITYCODE. Because one column is the description of the other column. Because the purpose of the analysis it to predict the accident severity, we will remove parameters that describe the damage after the accidents PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, HITPARKEDCAR. If we check the unique values in each column, PEDROWNOTGRNT, SPEEDING, INATTENTIONIND only has yes, no or nan. Thus, it would not be useful for prediction and removed. In addition, INCDATE and INCDTTM are reorganized to new columns called DOW, HOUR, and MONTH to present the day of week, hour of a day, and day of month. Because the large amount data we have, rows with nan and empty values are removed. For other columns, all categorical number will be adjusted to numerical numbers for further model training.

Reference:

[1]50+ Car Accident Statistics in the U.S. & Worldwide (Source: https://www.thewanderingrv.com/car-accident-statistics/)