Unterstanding Data

Data cleaning

There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. Thedataset has total observationsof 194673 with variation in number of observations for every feature. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there. These columns included pedestrian granted way or not, segment lane key, cross walk key and hit parked car.

There are many problems with the data set, considering that this is a machine learning project that uses classification to predict a categorical variable. The data set contains all observations of 194,673, with the number of observations varying for each attribute. First, the total data set was a high variation in the length of almost every column of the data set. The data set contained many empty columns, which unfortunately would have given us another advantage if the data had been available there. These columns included pedestrians with or without right of way, segment lane keys, cross lane keys and hit parked cars.

The goal of the model is to predict the severity of an accident, since the variable of severity was coded as 1 (only property damage) and 2 (injury collision) in the form of 0 (only property damage) and 1 (injury collision).

In addition, the Y was given the value 1, while Nan and no value 0 was given for the variables inattention, speed and under influence.

For the light status Light was given as 0, Medium as 1 and Dark as 2, for the road status Dry as 0, Mushy as 1 and Wet as 2. For the weather condition 0 is clear, overcast 1, windy 2 and rain and snow 3. 0 was assigned to the element of each variable that can be the least likely cause of a serious accident, while a high number represents an unfavorable condition that can lead to a higher accident severity.

While there were unique values for each variable, which were either "Other" or "Unknown", the complete deletion of these lines would have resulted in a large loss of data, which is not desirable.

To solve the problem of columns with different frequencies, arrays were created for each column, which were coded according to the original column and had the same proportion of elements as the original column. Then the arrays were imposed on the original columns at the positions with the elements "Other" and "Unknown". This entire process of data cleansing resulted in the loss of nearly 5000 rows containing redundant data, while other rows were filled with unknown values earlier.

Feature Section

Then, I began choosing columns to use from the dataframe that I created. The columns that I chose were SEVERITYCODE, which assigns a crash a value of 1, which means no injury, and 2, indicating injury, COLLISIONTYPE, which describes the type of crash, WEATHER, which describes the weather at the time of crash, ROADCOND, which describes the condition of the road at the time of crash, LIGHTCOND, which describes the light conditions at the time of crash, INATTENTIONIND, which describes whether the driver was distracted, and UNDERINFL, which describes whether the driver was under the influence.