IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Car Accident Severity

Help the Indian Government to prevent accidents...

2. Data Description

2.1 Data Source

The data come from “Kaggle” the world's largest data science community and also offers public data platform, where they have been published.

The dataset comprises of two csv files and one xls file:

Accident_train.csv: the training dataset

Accident_test.csv: the test dataset

AttributeLevelsDescription.xls

The dataset contains 8849 records and 17 columns in Accident train and 1549 records and 17 columns in Accident_test, which includes the data attributes such as Collision Reference Number, Policing area, Collision Severity etc... For good description of each attribute, you can refer to the Attribute Levels Description. Some or all can be used to train the model.

2.2 Data Pre-processing

To solve the problem our dataset should have balanced labels. In Data pre-processing, the data set is imputed by replacing NAN and missing values with most frequent values or median of the corresponding columns. All categorical variables are labeled by integers using Label encoder for each column.

The data is visualized for correlation. Negatively correlated features are selected to be dropped. Feature importance is plotted to visualize and only features with high importance are taken into consideration for accident severity.

2.3 Feature Selection

After data pre-processing, simply by reading the definition of each feature from the pdf mentioned above, we can see that there are some features whose values are derived from other features. There are mixed types of data, drop few columns due to its inconsistency. For selecting the best features, functions from sklearn library is used.