Contents

Prediction of Accident Severity & Pattern Extraction

Problem Description

Accidents in traffic lead to associated fatalities and economic losses every year worldwide and thus is an area of primary concern to Society from loss prevention point of view. Modelling accident severity prediction and improving the model are critical to the effective performance of road traffic systems for improved safety. In accident severity modelling, the input vectors are the characteristics of the accident, such as driver behaviour and attributes of vehicle, highway and environment characteristics while the output vector is the corresponding class of accident severity.

There are two main engineering approaches for dealing with traffic safety problems: the reactive approach and the proactive approach. The reactive approach, or retrofit approach, consists of making the necessary improvements to variable, for instance, existing hazardous sites in order to reduce collision frequency and severity at these sites. The proactive approach, on the other hand, includes a collision prevention approach, like, preventing a potential unsafe road conditions from occurring in the first place.

We focus on proactive approach which involves prediction of accident severity and working backwards, the concerned entity implements appropriate remedial measures to improve road safety.

By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities like Police, R&B and Transport from public policy point of view. The results of analysis and modelling can be used by these Departments to take appropriate measures to reduce accident impact and thereby improve traffic safety. It is also useful to the Insurers in terms of reduced claims and better underwriting as well as rate making.

The datasets are provided as cited below for the accident severity analysis

1. AttributeLevelsDescription_PHD.xls - depicts attributes description
2. Collision Data
    a. train_PHD.csv
    b. validation_PHD.csv
    c. test_NoTarget_PHD.csv
3. Base_Vehicle_Data_PHD.csv
4. NumberOfvehiclesbyCollosion_PHD.csv

Submission 1 (16th Sep 2017 EOD) :

Report on Exploratory Data Analysis, pre-processing, problem understanding in R Notebook or Jupiter notebook format (Use only train_PHD.csv for these tasks)

Submission 2 (17th Sep 2017 EOD) :

1. ***The details of Collision related datasets are mentioned below for analysis for First level submission:***
    a. Target attribute is "Collision Severity" for predictions.

b. "train_PHD.csv" & "validation_PHD.csv" are related to collisions/accidents.
   i. These datasets have the details depicting the circumstances of the collisions ie., details include the collision severity, day/date details, time, location, weather and road conditions, and carriageway hazards;
   ii. "train_PHD.csv" file should be used for visualization and also for modelling
   iii. "validation_PHD.csv" file should be used for testing your model performance.
c. "test_NoTarget_PHD.csv" should be used for evaluation of model.
   i. This is totally unseen data and hence it does not have the target attribute "Collision Severity".
   ii. The predictions obtained for this dataset should be uploaded to Kaggle.

2. The following Vehicle related datasets along with the data "train_PHD.csv" cited above should be considered for further analysis and pattern/rules extraction for second level submission:

a. "Base_Vehicle_Data_PHD.csv" gives the details of vehicles involved in each collision ie., the details include vehicle type, manoeuvre at the time of collision, and data about the driver (age, sex) etc.
   i. **Consider the subset from vehicles data based on the collisions that are there in train_PHD.csv only for patterns.**
   ii. This subset should be aggregated at collision-level with "Collision Reference No." as key while dropping the "vehicle reference" attribute and should be appended to "train_PHD.csv"
   iii. **Approach to prepare the aggregated data**: Each attribute can be made dummy having the number of vehicles corresponding to each level of that attribute. After the aggregation, the data may be further updated to binary for the purpose of pattern extraction.

For ex: Original Data

Collision Reference No.	Vehicle Reference No	Vehicle Type
1	1	2
1	2	8
1	3	8
1	4	3

Aggregated data would be :

Collision Reference No.	Vehicle Type 2	Vehicle Type 3	Vehicle Type 8
1	1	1	2

Note: You are always welcome to come-up with your own approach. Needless to add that the patterns should be meaningful.

b. "NumberOfvehiclesbyCollosion_PHD.csv" has the collision id and number of vehicles involved in each accident/collision. Add this Number of vehicles attribute also to the dataset for patterns.
c. You have to reduce the number of levels <= 5 in each attribute for extracting the patterns based on the details provided for the attribute levels.

Note : Reducing the number of levels in each attribute and using the important attributes for pattern extraction may help you to get good patterns.

3. Benchmark for prediction:

a. Accuracy : 80%
b. Recall for the level "1" : 40%

4. Instruction for pattern extraction:

a. Extract top 5 meaningful patterns for each “Collision Severity” level based on the best metrics of confidence & support

5. Documents to be submitted into piazza are cited below:

a. Commented code of Modelling with explanation why you have chosen your model as the best model.
b. Commented code of pattern extraction as mentioned in 4(a)

Submission 3 (22nd Sep 2017 EOD) :

- Final documentation with additional efforts and improvements made during the week using the prescribed template.
- Final commented code
- Final presentation for viva