Predicting the severity of an accident: Data acquisition

Alessandro Sgarabotto

12/10/2020

1 Data acquisition

1.1 Data sources

The data is provided as course material at this link. The dataset refers to severity of accidents occurred in the city of Seattle (US) from 2004 to present time. All collisions were provided by Seattle Police Department (SPD) and recorded by Traffic Records. The dataset has 194 673 observations and 38 features containing both numerical and categorical data (Table 1).

The first column is SEVERITYCODE and pertains to the severity code, namely the accident severity which is labeled as 3, in case of fatality, 2*b*, in case of serious injury, 1, in case of damage or 0, in case of unknown details. In the present dataset the severity code is a binary variable that is labeled as 1 for low severe accident or 2 for high severe accident. The objective of this project is the prediction of the severity code (i.e., the accident severity) by means of supervised machine learning algorithms. The other columns contain details about the accident such as:

- accident/report identifier (OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS);
- accident location (X,Y, INTKEY, LOCATION, SEGLANEKEY, CROSSWALKKEY);
- accident date and time (INCDATE, INCDTTM);
- type of collision (ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE, ST_COLDESC);
- people and vehicles involved in the accident (PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, PEDROWNOTGRNT, HITPARKEDCAR);
- road condition (ROADCOND);
- light condition (LIGHTCOND);

Feature	***Description***
SEVERITYCODE	A code that corresponds to the severity of the collision
X, Y	Latitude and longitude of the accident location
OBJECTID	Unique identifier
INCKEY	Unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Report number
STATUS	Report status
ADDRTYPE	General type of collision
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	
EXCEPTRSNDESC	
SEVERITYCODE.1	A detailed description of the severity of the collision
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
INCDATE	The date of the incident
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
SDOT_COLDESC	A description of the collision corresponding to the collision code
INATTENTIONIND	Whether or not a driver involved was under the influence of drugs or alcohol (Y/N)
UNDERINFL	A detailed description of the severity of the collision
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)
ST_COLCODE	A code provided by the state that describes the collision
ST_COLDESC	A description that corresponds to the stateŠs coding designation
SEGLANEKEY	A key for the lane segment in which the collision occurred
CROSSWALKKEY	A key for the crosswalk at which the collision occurred
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N)

Table 1: Dataset features. The feature SEVERITYCODE colored in red is the target feature.

- car speeding (SPEEDING);
- influence of drug/alcohol (UNDERINFL);
- lack of attention (INATTENTIONIND);
- further detailed description of the accident severity (EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC);

Not all the features are significant and negligible attributes for the aim of this investigation. Therefore the data has to be cleaned and wrangled before setting up a machine learning model.