Data Part

Car Accident Severity Data

2. Data acquisition and cleaning

Data

In this section the data that will be used to solve the problem is described. It contains the explanation, why the is date is adequate for the problem and is used. In the discussion part, examples of data is provided.

2.1 Data sources

To deal with accidents data the shared data for Seattle city is used as an example:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Metadata descriptions:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

2.2 Data cleaning

After open the CSV file and check what type of data is contained. The first column colored in
yellow is the labeled data. The remaining columns have different types of attributes. A selection of these attributes are used to train the model. Also most of the observations found are of sufficent quality to train and test the machine learning model to be build in this case.

The label for the data set used is severity, which describes the
fatality of an accident. It shall be notice that the shared data has unbalanced labels. So the data is balanced, otherwise, a biased ML model would be created.

2.3 Feature selection

The following is a list of attributes or features
that are used. For description of each attribute, it is refered to the web
link on the CSV file.

SEVERITYCODE,

STATUS,

ADDRTYPE - Collision address type:
- Alley
- Block
- Intersection

INTKEY, - Key that corresponds to the intersection associated with a collision

LOCATION, - Description of the general location of the collision

SEVERITYCODE, - A code that corresponds to the severity of the collision:
- 3—fatality
- 2b—serious injury
- 2—injury
- 1—prop damage
- 0—unknown

PERSONCOUNT, - The total number of people involved in the collision

PEDCOUNT, - The number of pedestrians involved in the collision. This is entered by the state.

PEDCYLCOUNT, - The number of bicycles involved in the collision. This is entered by the state.

VEHCOUNT, - The number of vehicles involved in the collision. This is entered by the state.

INCDATE, - The date of the incident.

INCDTTM, - The date and time of the incident.

JUNCTIONTYPE, - Category of junction at which collision took place

INATTENTIONIND, - Whether or not collision was due to inattention. (Y/N)

UNDERINFL, - Whether or not a driver involved was under the influence of drugs or alcohol.

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)

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. For more information about these codes, please see the State Collision Code Dictionary. Codes: 0-5, 10-32, 40-57, 60-67, 71-74, 81-84.

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)

In addition, there is probably need to do some feature engineering to improve the predictability
of the model as follows:

From the ST_COLCODE, a smaller set of cathegories could be defined.

From INCDATE the day of the week could be calculated.

From INCDTTM the part of the day could be calculated: morning, afternoon, evening, night.

The target or label columns should be accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. These terms are cathegories and are constructed from the last attributes listed above.

Then, the built severity data set is applied building a machine learning model.