**Data**

**Data source**

Collisions - Seattle GeoData – ArcGIS Online:

<https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0>

Attribute Information:

<https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>  
GeoJService:

[https://gisdata.seattle.gov/server/rest/services/SDOT/SDOT\_Collisions/MapServer/0/query?outFields=\*&where=1%3D1](https://gisdata.seattle.gov/server/rest/services/SDOT/SDOT_Collisions/MapServer/0/query?outFields=*&where=1%3D1)

GeoJSON:

<https://opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.geojson>

The source includes all types of collisions data in Seattle city from 2014/01/01 to 2020/05/20, There are 35 attributes and total 212453 records, the target dependency is the severity of the collision:

* 3—fatality (325)
* 2b—serious injury (2950)
* 2—injury (55964)
* 1—prop damage (131672)
* 0—unknown (20668)

There are 194673 records in 4 categories excluding the missing information (0). To simply the issue, they were divided into 2 categories: **1-prop damage**(1) and **2-injury**(2,2b,3).  
The Geo-information, Latitude and Longitude of collision, is integrated into the master source, here is the example of the data:

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

**Data selection**

184920 out of 194673 valid records are selected by ignoring "Unmatched" values in "STATUS" column and "Not Enough Information, or Insufficient Location Information" values in "EXCEPTRSNDESC" column. In addition, the 3 missing values in "ADDRTYPE" column are also deleted. As a result, around 5% records cannot be used for the training and are removed from the datasets.

**Feature selection**

In the 37 features, there are 22 may contribute to accidental severity in certain way, and they have to be transformed into appropriate data format for further processing and exploratory. The table below summarized the treatment of different features

|  |  |
| --- | --- |
| **Format** | **Selected Features** |
| Binary | INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR |
| Float | X, Y |
| Date | INCDATE |
| Time | INCDTTM |
| Encode Categorical | ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND, ST\_COLCODE |
| Int | PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT |