**Car Accident Severity Prediction**

Aurindom Mahanti

1. **Data acquisition and cleaning**
   1. **Data collect**

This dataset is the example dataset of the Coursera capstone project, which can also be accessed by this link (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv).

* 1. **Data Cleaning and feature selection**

The given data needs good amount of cleaning before getting used for the purpose of fatality prediction. There are a lot of missing row values in many columns that needs to be handled properly. In column “Status” two values are given (matched and unmatched) showing the status of given data after being verified. So the data that were not matching the record i.e. the rows marked “unmatched” had to be removed first.

Now there are a lot of columns that are needed to be removed because they don’t seem to have any meaning towards severity of the accident. Latitude(**X**) and longitude(**Y**) of the location of accident doesn’t seem to make any meaning towards severity of the accident, and for the same reason they had to be removed. **OBJECTID** is ESRI unique identifier, **REPORTNO, INCKEY** and **COLDETKEY** are unique identifier of the incident. **EXCEPTRSNCODE** and **EXCEPTRSNDESC** have less entries. **SEVERITYDEC** is just the description for the code **SEVERITYCODE** so basically both are same. Because of convenience **SEVERITYCODE** was selected out of two. **SDOT\_COLCODE** and **SDOT\_COLDESC.** In the similar fashion **ST\_COLCODE**, **SEGLANEKEY** and **ST\_COLDESC** doesn’t seem to be useful. **INCDATE** and **INCDTTM** are data and date-time of the accident that will just be helpful to the police and not to us for our classification.

Now from the rest of the feature we have to select the best feature that can be good for our analysis. The feature set is selected by looking at the Pearson correlation coefficient or Chi-square coefficient. Here we have used Pearson correlation, and hence we convert the categorical values to variables of type int64, int32 or float.

COLLISION\_NUM -0.128429

WEATHER\_NUM -0.110431

VEHCOUNT -0.080807

LIGHTCOND\_NUM -0.061206

ROADCOND\_NUM -0.049446

LOCATION\_NUM -0.038109

SDOTCOLNUM 0.004528

INTKEY 0.005955

X 0.010369

Y 0.018359

OBJECTID 0.026527

INCKEY 0.029442

COLDETKEY 0.029465

SPEEDING\_NUM 0.038249

UNDERINFL 0.043761

INATTENTIONIND 0.044839

SEGLANEKEY 0.104412

PERSONCOUNT 0.129782

CROSSWALKKEY 0.175533

SDOT\_COLCODE 0.186127

PEDCYLCOUNT 0.215718

PEDCOUNT 0.248121

SEVERITYCODE.1 1.000000

SEVERITYCODE 1.000000

Looking at the above table we can conclude saying that the best features to consider can be the ones which have a coefficient <0.01

**LOCATION\_NUM, INTKEY and SDOTCOLNUM** are good feature variables but have too many values that makes it a bad data for our use.

Data like **SPEEDING\_NUM,UNDERINFL** and **INATTENTIONIND** have slightly lesser correlation but always have influence on such cases. Hence we consider them as well.

Target: **SEVERITYCODE**

Features selected for further steps:

* **VEHCOUNT**
* **LIGHTCOND\_NUM**
* **ROADCOND\_NUM**
* **WEATHER\_NUM**
* **COLLISION\_NUM**
* **UNDERINFL**
* **INATTENTIONIND**
* **SPEEDING\_NUM**