**Seattle Car** **Accident Severity**

1. **Introduction**

Car accidents or on road collisions and something we witness daily on the news. The vehicle count on road today is much larger than it used to be 10 years ago.

The predictive analysis performed here aims towards analyzing the “Severity” of the accident/collision based on road conditions, lighting conditions, area of collision, number of people involved and many more factors as such. Knowing the severity of any such collision before hand will lead to prevention and prompt action.

1. **Data**

All the collision data used in this analysis is taken from ArcGIS, which was provided by Seattle Police Department and recorded by traffic records. The data provided is that of collisions which took place in the city of Seattle, from year 2004 till present.

Mentioned below is list of features that was available in the raw data:

|  |  |
| --- | --- |
| **SEVERITYCODE** | Target Column (1: Property Damage Only Collision, 2: Injury Collision) |
| **SEVERITYCODE.1** | Copy of Target Column |
| **SEVERITYDESC** | Description of Target Column |
| SDOTCOLNUM | A number given to the collision by SDOT |
| JUNCTIONTYPE | Category of junction at which collision took place |
| X, Y | Coordinate of accident |
| LIGHTCOND | The light conditions during the collision |
| WEATHER | A description of the weather conditions during the time of the collision |
| ROADCOND | The condition of the road during the collision |
| ST\_COLCODE | A code provided by the state that describes the collision |
| ST\_COLDESC | A description that corresponds to the state's coding designation |
| COLLISIONTYPE | Collision type |
| UNDERINFL | Whether or not a driver involved was under the influence of drugs or alcohol |
| LOCATION | Description of the general location of the collision/address |
| ADDRTYPE | Collision address type (Alley, Block, Intersection) |
| SDOT\_COLCODE | A code given to the collision by SDOT |
| SDOT\_COLDESC | A description of the collision corresponding to the collision code |
| SEGLANEKEY | A key for the lane segment in which the collision occurred |
| CROSSWALKKEY | A key for the crosswalk at which the collision occurred |
| VEHCOUNT | The number of vehicles involved in the collision |
| INCDTTM | The date and time of the incident |
| INCDATE | The date of the incident |
| PEDCYLCOUNT | The number of bicycles involved in the collision |
| PEDCOUNT | The number of pedestrians involved in the collision |
| PERSONCOUNT | The total number of people involved in the collision |
| STATUS | (Matched, Unmatched) \*\*\* |
| REPORTNO | Report identifier |
| COLDETKEY | Secondary Key for incident |
| INCKEY | Unique key for incident |
| OBJECTID | ESRI unique identifier |
| HITPARKEDCAR | Whether or not the collision involved hitting a parked car |
| EXCEPTRSNCODE | \*\*\* |
| EXCEPTRSNDESC | \*\*\* |
| PEDROWNOTGRNT | Whether or not the pedestrian right of way was not granted |
| SPEEDING | Whether or not speeding was a factor in the collision |
| INATTENTIONIND | Whether or not collision was due to inattention |
| INTKEY | Key that corresponds to the intersection associated with a collision |

\*\*\* Column details not available.

There are in total 38 data columns in the dataset including the 3 target related columns. We will keep various aspects in mind while deciding the importance of a particular column or the transformation it may need before we feed it to the model.

Some of the given data columns are features related to or identifying a single particular accident, thus may not be very much useful for our predictive analysis. These features include:

*SDOTCOLNUM, Coordinates, LOCATION, INCDTTM, INCDATE, REPORTNO, COLDETKEY, INCKEY, OBJECTID.*

There are some description columns for a given code. Columns *ST\_COLDESC, SDOT\_COLDESC and EXCEPTRSNDESC* are description columns for code which is already specified in the given dataset.

There are also data columns which has missing data in abundance. *Column EXCEPTRSNCODE,* EXCEPTRSNDESC, *PEDROWNOTGRNT, SPEEDING, INATTENTIONIND* and *INTKEY* have more than 50% of data missing. Although few of these columns can be very crucial indicator of collision severity, it would be misguiding to use it with so many missing rows and very difficult to fill in these categorical values.

Columns mentioned in all the three categories above will not be used in the model that we are going to build. Most of the columns that remains are categorical and will require one-hot and label encoding before we can use them as a feature for our model.