**Data**

The dataset contains 194,673 rows that contain historical data of prior accidents. There are 37 attributes in the dataset. The dependent variable is a column called SEVERITYCODE, which indicates the severity of the accident. The valid values for SEVERITYCODE are:

3 – fatality

2b – serious injury

2 - injury

1 – prop damage

0 - unknown

Several of the columns provide are in text format, which would make it difficult for machine learning. The columns that seem appropriate for machine learning are the numeric columns and the text columns that are codes or indicators. Some of the columns are inconsistent, having both numeric and charter data. The column UNDERINFL, which is whether or not a driver involved was under the influence of drugs or alcohol, has some rows contain N, some rows contain 0, and some rows contain 1.

Attributes that are valid for this analysis

|  |  |
| --- | --- |
| 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 |
| INJURIES | The number of total injuries in the collision. This is entered by the state |
| SERIOUSINJURIES | The number of serious injuries in the collision. This is entered by the state |
| FATALITIES | The number of fatalities in the collision. This is entered by the state |
| SDOT\_COLCODE | A code given to te collision by SDOT |
| INATTENTIONIND | Whether or not collision was due to inattention (Y/N) |
| ROADCOND | The condition of the road 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. |
| HITPARKEDCAR | Whether or not the collision involved hitting a parked car (Y/N) |

Since the dependent variable, SEVERITYCODE can have four different values, to support the business question, I may combine SERVERITYCODEs 3, 2b, and 2 into one code (1), and SERVERITYCODEs 1 and 0 into another code (0). Using a binary option, I would execute the logistic regression model.