**Predicting the Severity of Road Accidents**

# **Shilpa Banerjee**

## September 23,2020

1. **Introduction**
   1. **Background**

The Seattle community has expressed concern over the serious road accidents taking place in certain road junctions. Incidents of vehicular collisions as well as vehicles hitting pedestrians or cyclists have been reported.

The increasing number of serious accidents will eventually bring down the road safety score of this suburb there being leading to unpleasant consequences. Traffic jams caused by accidents are a nuisance for all those on the road.

* 1. **Problem**

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway.

Long lines of cars barely moving. As you keep driving, police car start appearing from

afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you,

given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

* 1. **Interest**

It would be great if there is something in place that could warn people, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be.

1. **Data acquisition and cleaning**
   1. **Data sources**

**I have used the dataset provided in the example. This is the Seattle accidents database which discusses the fatality of accidents based on a set of conditions. The fatality is measured using attribute SEVERITYCODE has been given values 1 or 2.**

|  |  |
| --- | --- |
| **SEVERITYCODE** | **Description** |
| **1** | **Property damage** |
| **2** | **Injury** |

**This score is based on several factors some of which are mentioned below.**

|  |  |
| --- | --- |
| **Location** | **Road Condition** |
| **Weather Condition** | **Junction** |
| **Car Speeding** | **Number of people involved** |
| **Light Conditions** | **Number of vehicles involved** |

**There is a total of 37 attributes which are a mix of numerical and categorical datatypes.**

* 1. **Data Cleaning**

**This data ranges from years 2004 to 2020 so it can be considered up to date for developing the model. There are however quite a number records with null values.**

**Out of 37 columns we can see redundant columns as we have categorical codes and their respective descriptions (SEVERITYDESC, SDOT\_COLDESC, ST\_COLDESC, LOCATION). These description columns have been removed.**

**There are a few ID columns (OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, SDOTCOLNUM, ST\_COLCODE, SEGLANEKEY, CROSSWALKKEY) as well which are not useful for our purpose and removed from the dataset.**

**The columns (INATTENTIONIND, UNDERINFL, EXCEPTRSNCODE, EXCEPTRSNDESC, PEDROWNOTGRNT, SPEEDING) has totally or mostly null values hence they are removed.**

**Some columns (INCDTTM,** **JUNCTIONTYPE) have redundant information and hence have been excluded.**

**Original dataset comprised of 37 columns and 194673 records. After cleaning up we have 13 columns and 184146 records remaining.**

* 1. **Feature Selection**

**The column SEVERITYCODE is our y value which has to be predicted.**

The columns WEATHER, ROADCOND and LIGHTCOND are closely correlated. Hence, I have considered only one out of these three for the model.

**HITPARKEDCAR is correlated to COLLISIONTYPE as we are indicating the same in that as well.**

Final Feature selection is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Used in Model** | **Description** | **Reason for Exclusion** |
| **ADDRTYPE** | Included | This describes the junction where accident occurred (Alley,Block,Intersection) |  |
| **COLLISIONTYPE** | Included | This describes the collision (Angle,Head on, Side swipe, Rear end, Parked car etc) |  |
| **PERSONCOUNT** | Included | The total number of people involved in the collision |  |
| **PEDCOUNT** | Included | The number of pedestrians involved in the collision |  |
| **PEDCYCLCOUNT** | Included | The number of bicycles involved in the collision. |  |
| **VEHCOUNT** | Included | The number of vehicles involved in the collision. |  |
| **WEATHER** | Included | A description of the weather conditions during the time of the collision. |  |
| ***HITPARKEDCAR*** | *Excluded* | *Whether or not the collision involved hitting a parked car* | *Correlated to COLLISIONTYPE* |
| ***ROADCOND*** | *Excluded* | *The condition of the road during the collision.* | *Correlated to WEATHER* |
| ***LIGHTCOND*** | *Excluded* | *The light conditions during the collision.* | *Correlated to WEATHER* |
| ***X*** | *Excluded* | ***X – coordinate of the collision location*** | *Not useful for model, may use for map* |
| ***Y*** | *Excluded* | ***X – coordinate of the collision location*** | *Not useful for model, may use for map* |
| ***INCDATE*** | *Excluded* | *Date of Incident* | *Not useful for model, used for studying spread of data.* |