

Introduction:

Road accidents are one of the biggest problems in any nation, it leads to property damage or some cases loss of life as well. Many families in world are suffering because of road accidents.

In order to reduce the frequency of Car accidents in a community, an algorithm needs to be developed based on previous accident incidents that occurred in past.

Currently I'm trying to build an algorithm based on Seattle city accident data from police department showing all the accidents occurred from 2004 till present. When conditions are bad this algorithm model will alert the drivers to be careful.

This analysis will be helpful and have many real-world applications, this data analysis will be useful for the road transport authority to improve conditions of roads and to mitigate accidents in bad weather situations.

Data:

The data from Seattle PD consists of more than 190,000+ incidents which have been collected over past 15+ years. This data set has high variation in almost every column of dataset. The dataset has lot of empty columns, it could have been beneficial if the data has been present in it.

The model aim is to predict the severity of an accident, considering that the variable of severity code is in the form of 1 (Property Damage only) and 2 (Injury Collision) ,which I changed to 0 and 1 respectively for purpose of ease of representation. Changed the string values in other parameters to int, for light condition, Light was given 0 along with medium as 1 and Dark as 2. For Road Conditions, Dry as 0, Mushy as 1 and wet as 2. As for weather condition Clear -0, Overcast-1, windy-2, Rain/Snow – 3 was assigned to each variable.

Feature Variables	Description
INATTENTIONIND	Whether the driver was inattentive? (Y/N)
UNDERINFL	Whether the driver was under alchohal infulence? (Y/N)
WEATHER	Weather Condition during collison (clear, rainy,..etc)
ROADCOND	Weather Condition during collison (wet, dry,..etc)
LIGHTCOND	Light Conditions during collisions (Lights on, Dark with lights on.)
SPEEDING	Whether the car is over the speed limit or not

The rows where the data is not proper and not complete , has been rremoved from the data frame. The is cleaning of data lead to almost loss of 5000 rows which had redudent data.

The Variables in above table is going to be our data for the algorithm model.

Number of entries in data for each variable - Seattle, Washington

