Predicting Severity of Collisions in Seattle

Author: Saurabh Patel

AUGUST 23, 2020

Contents

1. Introduction

Traffic collisions continue to be a serious problem. Roads safety is pressing concern for many countries, where road crash fatalities and disabilities is gradually being recognized as a major public health concern. According to World Health Organization (WHO); nearly 1.25 million people die in road crashes each year, on average 3,287 deaths a day. In addition, road traffic crashes rank as the 9th leading cause of death and account for 2.2% of all deaths globally.

Collisions are financial burden on government and society. Prediction of severity of collision helps local transport authority and emergency responders to manage traffic and avoid loss of life and property.

This project uses collision data of Seattle, WA. The aim of this project is to use data science methodology and machine learning to gain an understanding of the problem and predict the severity of collision and develop prevention mechanisms the same.

2. Data

2.1 Data Source

For this exercise we are using the data provided by released by the Washington State Dept. of Transportation (WSDOT). The dataset was hosted by coursera as part of the Data Science course. To understand the data, a supplementary metadata was also provided.

2.2 Data Cleaning

After importing dataset into Jupiter Notebook, a quick analysis showed data with missing values. From definition of the columns many of these attributes like speeding, inattention indictor could be used in prediction so it's better to clean such attributes.

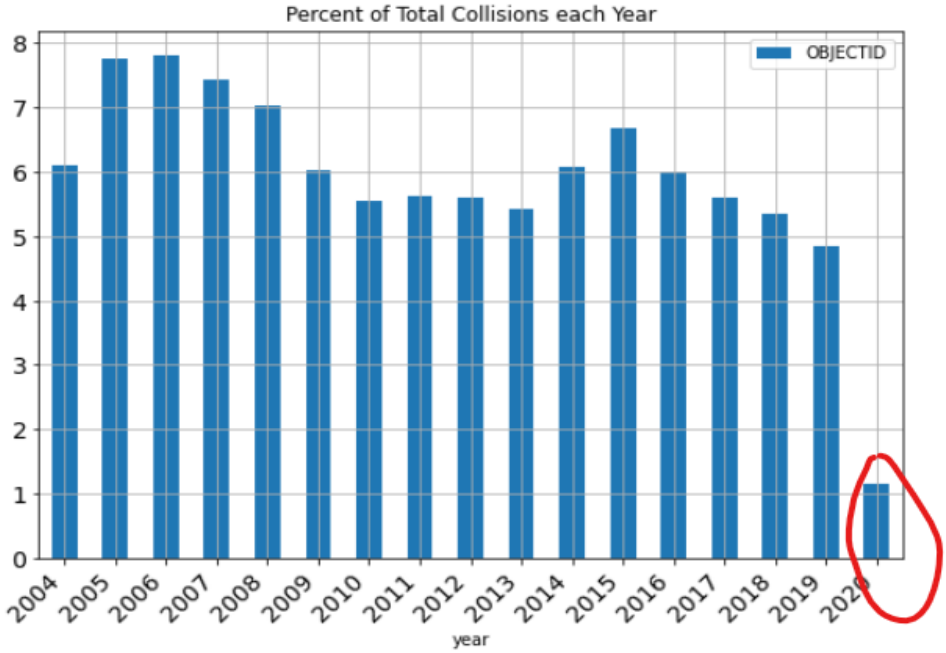
	data_type	percent_missing_values	total_unique_values
PEDROWNOTGRNT	object	97.60%	1
SPEEDING	object	95.21%	1
INATTENTIONIND	object	84.69%	1
JUNCTIONTYPE	object	32.50%	7
Y	float64	27.40%	23839
X	float64	27.40%	23563
LIGHTCOND	object	26.60%	9
WEATHER	object	26.10%	11
ROADCOND	object	25.70%	9
COLLISIONTYPE	object	25.20%	10
UNDERINFL	object	25.10%	4
LOCATION	object	13.80%	24102

The top 3 attributes can have possible values of Y & N as per the metadata. These attributes had only 'Y' in them to it was assumed that the null data was N. As prediction model use int the Y & N(null) were replaced with 1 & 0 respectively.

Based on metadata it is quite evident that many of the attributes are indicators (Y or N), a similar operation was performed on these attributes.

Col Name	Description	Values	Clean-up Action
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int
SPEEDING	Whether or not speeding was a factor in the collision.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int
INATTENTIONIND	Whether or not collision was due to inattention.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.	Y/N/0/1	As data contained multiple parameters, Y,N,1,0 it was streamlined to 1 & 0
HITPARKEDCAR	Whether or not the collision involved hitting a parked car.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int

The data spanned across from 2004 to 2020. For accurate prediction it is very necessary to remove noise from the data or inaccurate data. A simple bar chart shows data is distributed across year by percentage. It is quite evident that data from 2020 is not sufficient to train the model, hence data from 2020 was removed from the dataset.



For categorical data, null values were replaced with "Unknown", Other values were classified or binned to for a group. For example LIGHTCOND, day light was unchanged however Dark Condition were replaced with "Dark" and Dawn and Dusk with "Low Light". Such classification will help in fitting the model correctly. The table below shows transformation for categorical values.

Col Name	Clean-up Action
ADDRTYPE	Replace Null with "Unknown"
COLLISIONTYPE	Replace Null with "Unknown"
JUNCTIONTYPE	Mid-Block (not related to intersection) ->Non Intersection At Intersection (intersection related) -> Intersection Related Mid-Block (but intersection related) -> Intersection Related At Intersection (but not related to intersection) -> Non Intersection
LIGHTCOND	Dark - Street Lights On -> Dark Dark - No Street Lights -> Dark Dark - Street Lights Off -> Dark Dark - Unknown Lighting -> Dark Dusk -> Low Light Dawn -> Low Light
WEATHER	Blowing Sand/Dirt -> Not Clear Overcast -> Not Clear Raining -> Not Clear Severe Crosswind -> Not Clear Sleet/Hail/Freezing Rain -> Not Clear Snowing -> Not Clear Fog/Smog/Smoke -> Not Clear
ROADCOND	Wet -> Not Dry Ice -> Not Dry Oil -> Not Dry Sand/Mud/Dirt -> Not Dry Standing Water -> Not Dry Snow/Slush -> Not Dry