Car Accident Severity Prediction

Sudipto Ghosh

September 13, 2020

1 Introduction

1.1 Background

The Open Data Program makes the data generated by the City of Seattle has been openly available to the public for the purpose of increasing the quality of life for the residents, increasing transparency, accountability and comparability, promoting economic development and research, and improving internal performance management.

The Traffic Records Group, Traffic Management Division, Seattle Department of Transportation, provides data for all collisions and crashes that have occurred in the state from 2004 to the present day. The data is updated weekly and can be found at the Seattle Open GeoData Portal[1].

1.2 Motivation

We can exploit this data to extract vital features that would enable us to end up with a good model that would enable the prediction of the severity of future accidents that take place in the state. This would further enable the Department of Transportation to prioritise their SOPs and channel their energy to ensure that fewer fatalities result in automobile collisions.

2 Data

2.1 Data Understanding

The dataset is available as comma-separated values (CSV) files, KML files, and ESRI shapefiles that can be downloaded from the Seattle Open GeoData Portal. The data is also available from RESTful API services in formats such as GeoJSON. We download the dataset to our project directory and take a look at the data types and the dimensionality of the data. We could see that the dataset contains 221,389 records and 40 fields.

[1] https://data-seattlecitygis.opendata.arcgis.com/

The metadata of the dataset can be found from the website of the Seattle Department of Transportation[2].

On reading the dataset summary, we can determine the description of each of the fields and their possible values. The data contains several categorical fields and corresponding descriptions which could help us in further analysis. We made an attempt at understanding the data in terms of the fields that we shall take into account for later stages of model building.

As the dataset has possibly been sourced from a database table, several unique identifiers and spatial features are present in the database which may be irrelevant in further statistical analysis. These fields are OBJECTID, INCKEY, COLDETKEY, INTKEY, SEGLANEKEY, CROSSWALKKEY, and REPORTNO. Other fields such as EXCEPTRSNCODE, SDOT_COLCODE, SDOTCOLNUM and LOCATION and their corresponding descriptions (if any) are categorical but have a large number of distinct values that shall not be that much useful for analysis. The INCDATE and INCDTTM denote the date and the time of the incident but may not be of use in further analyses. The data needs to be pre-processed.

2.2 Statistical Insights

3 Methodology

4 Results

5 Conclusion

6 Future Work

[2]https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Field Name	Description
X	Longitude (in degree decimal)
Y	Latitude (in degree decimal)
OBJECTID	ESRI unique identifier
INCKEY	Unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	NA
STATUS	NA
ADDRTYPE	Collision address type: [Alley, Block, Intersection]
INTKEY	Key to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	NA
EXCEPTRSNDESC	NA
SEVERITYCODE	Code corresponding to the severity of the collision
SEVERITYDESC	Detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	Total number of people involved in the collision
PEDCOUNT	Number of pedestrians involved in the collision
PEDCYLCOUNT	Number of bicycles involved in the collision
VEHCOUNT	Number of vehicles involved in the collision
INJURIES	Number of total injuries in the collision
SERIOUSINJURIES	Number of serious injuries in the collision
FATALITIES	Number of fatalities in the collision
INCDATE	Date of the incident
INCDTTM	Time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	Code given to the collision by SDOT
SDOT_COLDESC	Description of the collision by SDOT
INATTENTIONIND	Whether or not collision was due to inattention
UNDERINFL	Whether it was under the influence of drugs or alcohol
WEATHER	Description of the weather conditions
ROADCOND	Condition of the road during the collision
LIGHTCOND	Light conditions during the collision
PEDROWNOTGRNT	Whether the pedestrian right of way was not granted
SDOTCOLNUM	Number given to the collision by SDOT
SPEEDING	Whether speeding was a factor in the collision
ST_COLCODE	Code provided by the state that describes the collision
ST_COLDESC	Description corresponding to the state's coding scheme
SEGLANEKEY	Key for the lane segment in which the collision occurred
CROSSWALKKEY	Key for the crosswalk at which the collision occurred
HITPARKEDCAR	Whether the collision involved hitting a parked car

Table 1: Fields in the Dataset