Project :

Accident Prevention Study based on Data Science Application on Car Collisions Data for City of Seattle.

Author: Andre K.

Date: September 2020

DISCLAIMER : This project is conducted for the purpose of completing capstone project for IBM Data Science Professional certification. Although the project makes use of publicly available car accident data in Seattle, there is no relation of this project to the city of Seattle and result of this project, as a whole or part, should not be used outside of the purpose of completing the above-mentioned capstone project.

I. INTRODUCTION

The authority of Seattle has been able to collect nearly 200,000 cases of car collisions over the span of 15+ years (2004-2020), which should serve as a great resource to understand why collisions happen, and most importantly, how can we prevent collisions to happen, in particular for the case of severity level 2, which is collision that cause injuries or fatalities. This project aims to understanding the attributes of collisions that can be drawn from raw data, to name a few: road condition, weather, and then applies scientific data analysis mechanism (i.e. machine learning) to model how those attributes contribute to causing collisions level 1 and 2. The result of the model is tested using appropriate methodology to ensure accuracy. Using the accident prediction model built from this project, based on the attributes being observed at any given time, traffic authorities can immediately take appropriate measures to prevent collisions.

II. DATA

This project utilizes the car collisions data made available by city of Seattle. The data spans over period from January 2004 to May 2020, which contains nearly 200,000 of car collisions data. The target variable (also known as “label”) of the data is “accident_severity_level”, which indicates whether the accident being observed is limited to property damage (level 1), or involves people injury or fatality (level 2). Other than the target variable, There are 37 attributes (columns), as summarized in following table:

Column Name	Description
SEVERITYCODE	The target variable, or label, where code 1 indicates property damage and code 2 indicates injury/fatality
X	Latitude coordinate of location
Y	Longitude coordinate of location
OBJECTID	Case Unique Identifier
INCKEY	Incident Key
COLDETKEY	Secondary Key
REPORTNO	Report Number
STATUS	Matched/Unmatched (for official reference purpose)
ADDRTYPE	Address Type (Alley, Block, Intersection)
INTKEY	Intersection Key (Applicable when Address Type is Intersection)
LOCATION	Address where collision is located
EXCEPTRSNCODE	Exception Code when location information is missing
EXCEPTRSNDESC	Exception Code description
SEVERITYCODE	Copy of Severity Code
SEVERITYDESC	Severity Code description
COLLISIONTYPE	Type of collision
PERSONCOUNT	Number of person involved
PEDCOUNT	Number of pedestrian involved
PEDCYLCOUNT	Number of cyclist involved
VEHCOUNT	Number of vehicle involved
INCDATE	Incident Date
INCDTTM	Incident Date and Time
JUNCTIONTYPE	Type of junction where collision happened
SDOT_COLCODE	Code use by SDOT (Seattle Department of Transportation)
SDOT_COLDESC	SDOT Code description
INATTENTIONIND	"Y" if collision is caused by Inattention
UNDERINFL	"Y" if collision is found related to drug or alcoholic influence
WEATHER	Weather condition at the time of collision
ROADCOND	Road condition at the time of collision
LIGHTCOND	Light condition at the time of collision
PEDROWNOTGRNT	"Y" if pedestrian right of way is not granted
SDOTCOLNUM	Collision number assigned by SDOT
SPEEDING	"Y" if speeding was involved
ST_COLCODE	Code used by State
ST_COLDESC	Description of code used by State
SEGLANEKEY	Key to indicate lane segment
CROSSWALKKEY	Key to indicate crosswalk
HITPARKEDCAR	"Y" if the collision involves hitting parked car

For data analysis purpose, we can drop some of the columns that are clearly not related to the attributes of collision, such as IDs. For the remaining data, we can further provide explanations on whether they can potentially be important attributes for collisions. Some of the attributes can also be grouped into same category.

Category	Column Name	Potential use as attributes
Target Variable	SEVERITYCODE	The target variable, or label, where code 1 indicates property damage and code 2 indicates injury/fatality
Location	X	Coordinate information can point out to location where collision is more frequent. Having said that, given the proximity of location, it can be hard to use coordinate to accurately allocate which location within city of Seattle that collision happens.
	Y	
	ADDRTYPE	Address Type (Alley, Block, Intersection). We can presume that collision is likely to happen at intersection, which we will study further in our analysis.
	LOCATION	Address where collision is located can be important attribute to indicate location where collision frequently happen, however given the large amount of different address, it can be hard to use address as useful attribute in our analysis
	JUNCTIONTYPE	Type of junction where collision happened can serve as important information, in particular if address type "intersection" is used as decision tree parameter
	SEGLANEKEY	Key to indicate lane segment. To use location information as attribute, this can be more practical to use than address.
	CROSSWALKKEY	Key to indicate crosswalk. To use location information as attribute, this can be more practical to use than address.
Type	COLLISIONTYPE	These columns indicate type of collision, which can be useful attribute related to collision and its severity. Further analysis is needed to potentially reduce the columns as the information can be redundant. For analysis, we can also remove the column that contains description, as for data processing purpose we only need the codes. For example: [Collision Type] : "Parked Car" is redundant with column [HITPARKEDCAR] so we can drop [HITPARKEDCAR] column. Similarly, [Collision Type] : "Angles" is redundant with [SDOT_COLCODE]: "11", [SDOT_COLDESC]: "MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE", as well as with [ST_COLCODE]: "10", [ST_COLDESC]: "Entering at angle". Presumably, other than [COLLISIONTYPE], we can drop other type related columns.
	SDOT_COLCODE	
	SDOT_COLDESC	
	ST_COLCODE	
	ST_COLDESC	
	HITPARKEDCAR	
Number involved	PERSONCOUNT	This information is important attribute as more number involves in collision, the more likeliness it will cause injury or fatility. Although we potentially can reduce the attributes by only keeping [PERSONCOUNT] and drop the other columns, pedestrians and cyclists are more prone to injury so our data analysis need to study if we need to keep each of these columns as attributes related to severity level.
	PEDCOUNT	
	PEDCYLCOUNT	
	VEHCOUNT	
Date / Time	INCDATE	Dates can provide information on holiday period or weekend where traffic behavior is different than weekday. Time is also important indication as we can analyze how the trafic condition at rush-hours, morning, afternoon, evening are related to collisions.
	INCDTTM	
Violation	PEDROWNOTGRNT	"Y" if pedestrian right of way is not granted
	SPEEDING	"Y" if speeding was involved
	INATTENTIONIND	"Y" if collision is caused by Inattention

	UNDERINFL	"Y" if collision is found related to drug or alcoholic influence
Environment condition	WEATHER	Weather condition (raining, snowing, severe crosswinds etc.) is likely important attribute to cause collision.
	ROADCOND	Road condition (dry, ice, wet, oil, sand etc.) is likely important attribute to cause collision.
	LIGHTCOND	Light condition (daylight, dawn, dark etc.) is likely important attribute to cause collision.

Having been able to identify and reduce the dimension of potential attributes to use in data analysis, next step is to examine quality and availability for each attributes, which include following aspects:

a. Unbalance (Skewness) of target variable.
   Data of target variable is skewed to collision severity level 1, which means there are significantly more data for severity level 1, compared to level 2. Therefore in our data analysis we need to "balance" the data using appropriate statistical methodology in our methodology section.
b. Missing data.
   If the attributes that we select do not have sufficient data, it maybe better to remove the attribute as it will impair predictability of the model to be built. We will analyse the data sufficiency in our section on methodology.
c. Data conversion.
   In order to be processed further for analysis. Attribute which has object type need to be converted to integer. For example: [COLLISIONTYPE] has description such as "Parked Car", "Rear Ended", which need to be converted into [1,2,…].
   Similarly, attribute with "Y" value, such as [SPEEDING] need to be converted into binary [0,1] for data processing and modelling purpose.

Understanding of the data as described above will pave a strong foundation to move into next section about methodology.