PREDICTING CAR ACCIDENT SEVERITY

Kim-Cuong Nguyen

Report Date: 09/15/2020

1. Introduction

1.1. Background

Car accidents and traffic jam have undeniably always been a headache problem to all car drivers. Car accidents happen for a lot of reasons. Those can stem from drivers' own mistakes as carelessness, alcohol, or speeding. Besides, there are also cases that the drivers get involved into collisions due to uncontrollable factors. For example, drivers are more prone to accidents when it is raining, and the road becomes wet and slippery. Or, even though the weather is clear, accidents are more likely to happen at several locations due to the characteristics of those places that make it more difficult for drivers to watch out for other drivers.

1.2. Objective

Due to the likelihood of car accidents to happen along the roads and the negative impacts they create, I was motivated to build a model to predict the possibility of a car accident and its severity. Specifically, I will use the data science powers to build a predictive model that can warn drivers about the possibility of getting into a car accident and the potential severity, so that the drivers can drive more carefully or reroute if possible.

2. Data

2.1. Data Overview

To meet the objective, I employed a dataset that lists collisions from 2004 to May 2020 in Seattle, recorded by Seattle Department of Transportation (SDOT). This is comprised of all types of collisions that display at the intersection or mid-block of a segment. The dataset is updated weekly.

The dataset is comprised 194,673 observations that are details on collisions during the timeframe. 38 variables are included, of which 18 are categorical variables with lots of them being high cardinality.

Appendix provides a brief explanation on the variables.

2.2. Data Cleaning

2.2.1. Irrelevant/Redundant Features

By examining the meaning of each variable within the dataset, I decided to remove the following ones and keep the meaningful variables only.

- **SEVERITYCODE:** A code that corresponds to the severity of the collision. This column is the same as column 'SEVERITYDESC' which gives the detailed description of the severity of the collision. Therefore, either of the two columns can be removed.
- **OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY:** These are identities assigned to the collisions by SDOT and have no meaning in predicting the accident; thus, can be removed.
- **STATUS**: There is no description available for this attribute. It has two values "Matched", and "Unmatched".
- **ADDRTYPE, LOCATION**: Detailed address of collision locations. This corresponds with the two columns of longitude and latitude. The columns of longitude and latitude are better inputs into the model because the location attribute is of high cardinality.
- **EXCEPTRSNCODE, EXCEPTRSNDESC:** The two columns gives some notes about data as "not enough information", etc.
- **COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR**: The attributes give details on collision outcomes, while the models aim at predicting the accident. Thus, the variables will be removed.
- **INCDATE**: The date of accident. This is part of another variable "**INCDTTM**" that shows date and time of the accident.
- **JUNCTIONTYPE**: This attribute is similar to "**ADDRTYPE**".

The sample observations of the dataset are as below:

X	Y	SEVERITY-DESC	INCDTTM	WEATHER	ROAD-COND	LIGHT-COND
-122.323	47.70314	Injury Collision	3/27/2013 14:54	Overcast	Wet	Daylight
-122.347	47.64717	Property Damage Only Collision	12/20/2006 18:55	Raining	Wet	Dark - Street Lights On

2.2.2. Missing Values

The number of observations with missing values is insignificant per variables. The total number of observations with missing values is 10,506, accounting for about 5.4%. Therefore, I decided to remove all missing values from the dataset.

X	Y	SEVERITY-DESC	INCDTTM	WEATHER	ROAD-COND	LIGHT-COND
2.74%	2.74%	0%	0%	2.61%	2.57%	2.66%

By further checking the values of three variables, WEATHER, ROADCOND, and LIGHTCOND, I found that the variables have the value 'Unknown' that is similar to missing values. The number of observations with this value is 17,462, accounting for 9.5% of the remaining observations, which is not significant. Therefore, I removed all these observations.

Because the original dataset is too large, leading to my failure to complete modelling due to a lack of computational capabilities and infrastructure resource, I decided to reduce the data size by using data from 2018 to present only, including 18,708 observations.

APPENDIX – Variable Definitions

X	Longitude of the collision location
Y	Latitude of the collision location
OBJECTID	ESRI unique identifier
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Another key for the incident
STATUS	N/A
ADDRTYPE	Collision address type: Alley, Block, Intersection
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	Code to indicate data status
EXCEPTRSNDESC	Description of the data status
SEVERITYCODE	A code that corresponds to the severity of the collision
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place.
SDOT_COLCODE	A code given to the collision by SDOT
SDOT_COLDESC	A description of the collision corresponding to the collision code.
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision.
ST_COLDESC	A description that corresponds to the state's coding designation.
SEGLANEKEY	A key for the lane segment in which the collision occurred.

CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)