Predicting the car accident rate

Yan Yan Choi

August 24, 2020

1. Introduction

1.1 Background

Car accidents happen frequently around the globe which contributes to a number of disastrous impacts on both human and property. However, many people cannot not deal with the root problem and make the same mistake repeatedly as they believe that car accident is often an unforeseen misfortune. In view of this, through leveraging the accident data, we would like to examine different accidents' severity, (i.e., under what kind of weather, will the car accident happen? are road and light condition the factor leading to the car accident?) which would certainly help us identify the culprit of car accident and build a better society.

1.2 Problem

Data is used to predict the severity of different accidents. We will mainly investigate the following question.

I.   Which location has the highest severity of accidents?
II. Under what weather conditions, will the accidents occur frequently?
III. Under which combination of factors, will severity of accident be the highest?

1.3 Target Audience

I.   Driver
When they know which route or location has the high accident rate, they can bypass the route and choose another path to arrive the same destination.

II. Passenger
They can select another public transportation when they can examine which day will be more likely to have car accidents.

III. Government
The research gives an insight to the government that under which conditions will the accidents most likely to happen, and thus government can implement a series of measures to tackle the problem, (i.e., apply IT techniques or monitoring system)

2. Data acquisition and cleaning

2.1 Data source

Data-Collision dataset that is provided by IBM will be used to examine the severity of accidents. This dataset incorporates of multiple rows and columns, including the weather condition, light condition and road condition on the day of accident.

2.2 Data cleaning

Some records are missing, such as light and weather conditions. In this case, the blank value will be replaced with not available, and will not be used in the analysis.
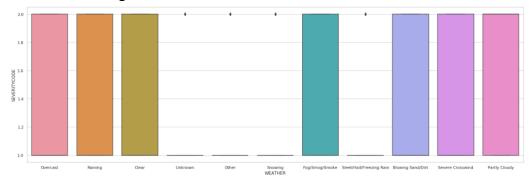
3. Exploratory Data Analysis

3.1 Relationship between weather condition and car accident's severity

From the table on the left, it is clearly seen that the car accidents most likely occur in clear day (111135/194673). Accidents commonly occur in raining and overcast day as well with 33145 and 27714 cases respectively.

	WEATHER
Clear	111135
Raining	33145
Overcast	27714
Unknown	15091
Snowing	907
Other	832
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5

From the figure below, accidents with higher severity will be more likely to happen in overcast, raining and clear day, which give us the insight that weather will be one of the factors contributing to the serious traffic accidents.

3.2 Relationship between road condition and accident

Regarding the road condition, it is crystal clear that dry road condition would be a major factor leading to the accident occurs. The above analysis also illustrates that clear day will lead to traffic accident. It is not hard to predict that accidents are more likely to occur under clear and dry conditions.

	ROADCOND
Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

3.3 Relationship between light condition and accident

The table on the left illustrates the relationship between light condition and accident. Daylight may easily lead to the accidents, which shows that accidents may be more likely to occur in the morning and afternoon.

	LIGHTCOND
Daylight	116137
Dark - Street Lights On	48507
Unknown	13473
Dusk	5902
Dawn	2502
Dark - No Street Lights	1537
Dark - Street Lights Off	1199
Other	235
Dark - Unknown Lighting	11

3.4 High severity with the most common combination of factors

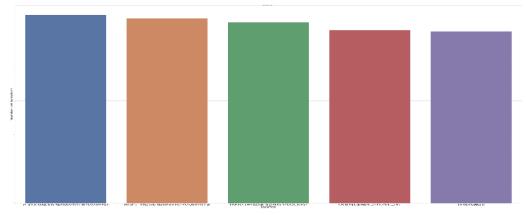|	WEATHER	ROADCOND	JUNCTIONTYPE	SPEEDING	ST_COLCODE	SEVERITYCODE
1	Blowing Sand/Dirt	Snow/Slush	Mid-Block (not related to intersection)	Y	50	2.0
11	Clear	Dry	At Intersection (but not related to intersection)	Y	13	2.0
18	Clear	Dry	At Intersection (intersection related)	Y	0	2.0
19	Clear	Dry	At Intersection (intersection related)	Y	1	2.0
28	Clear	Dry	At Intersection (intersection related)	Y	25	2.0
33	Clear	Dry	At Intersection (intersection related)	Y	30	2.0
35	Clear	Dry	At Intersection (intersection related)	Y	45	2.0
39	Clear	Dry	At Intersection (intersection related)	Y	73	2.0
41	Clear	Dry	At Intersection (intersection related)	Y	0	2.0
42	Clear	Dry	At Intersection (intersection related)	Y	1	2.0
49	Clear	Dry	At Intersection (intersection related)	Y	2	2.0
50	Clear	Dry	At Intersection (intersection related)	Y	21	2.0
53	Clear	Dry	At Intersection (intersection related)	Y	24	2.0

The above table describes which combination of conditions will lead to the higher severity of accidents. The top one is the combination of blowing sand/ dirt with snow will lead to the accidents occur easily. Besides, it can be clearly shown that speeding is also the culprit contributing to the car accidents.

Another common combination is clear weather with dry road condition. Serious traffic accidents are also likely to happen at intersection.
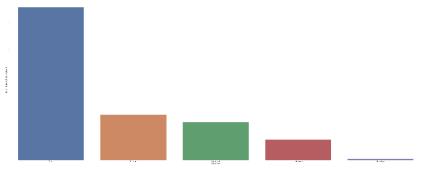
4. Modeling

Regression model and charts will be applied to predict the car accidents' severity. These will give further information on the reasons of an accident happened.
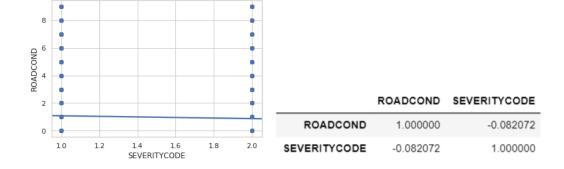
4.1 Location

From the figure, we can identify which location will have a higher chance of traffic accidents. The top one is Battery ST Tunnel NB Between Alaskan WY VI AND AURORA AVE N.

4.2 Weather

The figure shows that accidents usually occurs in clear weather, which is in accordance with the analysis in previous part.

4.3 Road condition

	ROADCOND	SEVERITYCODE
ROADCOND	1.000000	-0.082072
SEVERITYCODE	-0.082072	1.000000

From the above table, the correlation between road condition and severity is negative. And most of the serious accidents will occur under dry weather. Below is the table which converts categorial variable of road conditions to numerical values.

Road Condition	Number
Wet	0
Dry	1
Unknown	2
Snow/Slush	3
Sand/Mud/Dirt	4
Standing Water	5
NaN	6
Ice	7
Other	8
Oil	9

5. Conclusion

In this research, I analyze the correlation between different variables and the factors leading to the car accidents. From the finding, I identify that clear weather, dry road condition, and speeding would be the main culprit, as well as the accidents are more likely to occur at intersection. After understanding these, I believe that a number of measures can be implemented to tackle this problem (i.e., install monitoring system, warning signals and indicators in specific location). As such, the problem can be eased efficiently and effectively.

6. Future directions

Decision tree and classification model can be used to further study this research topic. Besides, training and testing data can also be used to verify the result. Moreover, since this study only focus on one dataset, I believe that more datasets should be considered if a more comprehensive study is needed. In the future research, it is hoped that more variables and conditions can be incorporated with different combinations.