Capstone Project - Car Accident Severity

Weihan Tang

September 5, 2020

Contents

1 Data Understanding and Preparation

1.1 Data Sources

In this study, we will use the dataset provided by the course of Applied Data Science Capstone project. This dataset contains nearly 200000 cases of car accidents, which is sufficient for our study.

1.2 Data Selection and Cleaning

1.2.1 Duplicate Columns

First of all, the target variable that we try to predict in this study is the severity code. This variable has two values, 1 and 2, whereas a value of 1 suggests property damage only, and a value of 2 indicates injury is involved. A glance at this dataset shows that there are two duplicate columns that contain data of this variable. Thus, we start by dropping one of the columns.

1.2.2 Feature Selection

This vast dataset contains information that is clearly not related to the cause of a road accident, such as OBJECTID, INCKEY, and etc. Here we only keep meaningful data that potentially have an impact on the cause of a road accidents, which include the following:

- Location of the accident: 'X', 'Y'
- Is the location a intersection or a block: 'ADDRTYPE'
- Weather condition: 'WEATHER'
- Under influence of achohol: 'UNDERINFL'
- Road condition: 'ROADCOND'
- Light condition: 'LIGHTCOND'item Speeding or not: 'SPEEDING'

1.2.3 Missing Values

The dataset contains many values of NaN. These values need to be dealt with in different ways.

- The location data 'X'and 'Y'have a few rows of missing data. In this case, the entire rows are dropped. The same approach is applied to the other features where the number of NaN values contributes to only a small portion of the whole columns.
- An exception is the data contained in 'SPEEDING'. Due to the data format contained in the raw spreadsheet, only a speeding case has an entry 'Y', the rest are missing values. In this case we replace the NaN values with 0.

1.2.4 Data Formatting and Filtering

Many features in the dataset has string values. To prepare this dataset for the downstream analysis, we need to replace the categorical values with quantitative values. Here we use one-hot binning to convert categorical data into a number of indicator columns. For example, the 'ROADCOND'column that contains the road condition data is converted into 9 separate columns - 'Dry',

'Ice', and etc, each containing binary values indicating whether an accident happened under such road conditions.

A further investigation shows that 'Dry'and 'Wet'conditions contribute to 91% of the total number of cases. To simplify the dataset, we only keep 'Dry'and 'Wet'columns, and drop the rest. The two columns are concatenated with the dataset, and the original column 'ROADCOND'is dropped. This approach is also applied to other columns 'ADDRTYPE', 'WEATHER', and 'LIGHTCOND'.

Another issue the data set has is that columns such as 'UNDERINFL'contains mixed values of Y, 0, and N. Both N and 0 indicate the driver is not influenced by achohol. Here we convert values of N to integer 0, and Y to integer 1. We apply the same approach to column 'SPEEDING'. Also, upon checking the data types of each column, we notice that column 'UNDERINFL'have a type of 'object'. We need to cast these two variables into integers for further analysis. After the above process, the data is now cleaned, formated, and filtered for modeling and downstream analysis, as shown below.

	SEVERITYCODE	X	Y	UNDERINFL	SPEEDING	Dry	Ice	Block	Intersection	Clear	Raining	Overcast	Daylight	Dark - Street Lights On
0	2	-122.323148	47.703140	0	0	0	0	0	1	0	0	1	1	0
1	1	-122.347294	47.647172	0	0	0	0	1	0	0	1	0	0	1
2	1	-122.334540	47.607871	0	0	1	0	1	0	0	0	1	1	0
3	1	-122.334803	47.604803	0	0	1	0	1	0	1	0	0	1	0
4	2	-122.306426	47.545739	0	0	0	0	0	1	0	1	0	1	0

Figure 1: Cleaned and Formatted Dataset