Predicting Road Accidents: An analysis of the main factors causing car accidents in Seattle, USA

Luis Terán

1. Introduction

a. **Background:** Motor vehicle accidents continue to be one of the leading causes of accidental deaths and injuries in the United States. They are responsible for billions in property damage and other economic losses each year. More than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. An additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of death in the U.S. for people aged 1-54 (ASIRT, 2020).

b. **Problem:** In order to identify the cause of the problem, the aim of this project is to identify those factors that influence the most on cars accidents and have a quantitatively estimate of the significance the relationship between the factors and the road accidents.

c. **Interest:** Even though this is a sample data from Seattle, this behavior patterns in car accidents can be related to other states or even countries with less amount of data of this problem. So this can be used as a feature reference for prevention of car accidents in all other places.

2. Data

a. **Acquisition:** The original dataset was obtained from the Seattle's government page at:
https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d
The dataset if available for public access. It was created in April 8, 2020 and last update register is from August 27, 2020. Further information of the dataset is available in:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

b. **Feature selection:** Before selecting the features for predicting it was necessary to state our main prediction objective. We want to know the probability and magnitude of an accident given some characteristics, in this way, the "Severity code" variable could be helpful. The severity code variable is the severity of the collision in a road accident, that is our expected to predict variable.
The complete dataset was split into two different dataframes. The first one was for feature selection, those variables that can contribute to an accident were selected and joined to this dataframe. The data description is shown as follows:

Attribute	Data type	Description
INCDATE	Date	Date of the accident
UNDERINFL	Boolean Values: (1: Yes, 0: No)	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	Int Values: 1: Clear 2: Raining 3: Overcast 4: Snowing 5: Fog/Smog/Smoke 6: Sleet/Hail/Freezing Rain 7: Blowing Sand/Dirt 8: Severe Crosswind 9: Partly Cloudy 10: Other/Unknown	A description of the weather conditions during the time of the collision.
ROADCOND	Int Values: 1: Dry 2: Wet 3: Ice 4: Snow/Slush 5: Standing water 6: Sand/Mud/Dirt 7: Oil 8: Other/Unknown	The condition of the road during the collisions
LIGHTCOND	Int Values: 1: Daylight 2: Dark – Street Lights On 3: Unknown 4: Dusk 5: Dawn 6: Dark – No Street Lights 7: Dark – Street Lights On 8: Dark – Unknown Lighting 9: Other	The light conditions during the collisions
SPEEDING	Int (1: Yes, 0: No)	Whether or not speeding was a factor in the collisions

On the other hand, there were selected other variables for additional information. The aim of this dataset is for getting a better understanding of the data and the problem.

c. **Data wrangling:** Since the amount of null observations in the data set was not significant for most of the variables, the null observations were deleted without taking into account the "SPEEDING" column. Then for every variable in the dataset, the data was transformed in order to get only integer values:
SEVERITYCODE: Transformed into a 4 integer level of severity,
UNDERINFL: Transformed combined Y/N and 1/0 observations into only 1/0.
WEATHER: Transformed categorical data into integers.
ROADCOND: Transformed categorical data into integers.
LIGHTCOND: Transformed categorical data into integers.
SPEEDING: It was assumed that the missing values refer to the opposite of filled values.

3. **Exploratory data analysis**
4. **Predictive Modeling**
5. **Conclusions**
6. **References**

- ASIRT, 2020 (Association for Safe International Road Travel): Annual United States Road Crash Statistics https://www.asirt.org/safe-travel/road-safety-facts/