Data Wrangling and Exploration on Toronto's Pedestrian collisions and Injured Persons Data

2023-01-09

1. Background

A bad data set that has errors is more likely to produce unreliable insights. This exercise is grouped into two main sections, which are data wrangling and data exploration. The goal is to produce a reliable dataset that can be used to produce reliable insights for the Ministry of Transportation.

1.1 Data Wrangling

In this section, a) I will familiarize myself with the data (data discovery), b) clean the data by removing potential outliers, drop missing/empty information or replace them using imputation if appropriate, and identify row and column duplicates, c) extract relevant columns or create new variables from existing columns, and d) validate the data by ensuring that columns contain consistent, accurate and expected range of values.

A. Data Discovery

The number of rows and columns of the collision data are 504,069 and 12, respectively. Eight of the variables are non-numeric. The involved_persons data, which focuses on pedestrian collisions, has 16,045 rows and 15 columns or variables. Most of the columns of both data have significant amounts of missing (NA) or empty information (Table 1 and 2).

Table 1. Number of missing and empty information in the collision data

variable	sum_missing	sum_empty
collision_no	0	0
accdate	0	NA
district	0	39738
latitude	13	NA
longitude	13	NA
impact_type	0	276
light	0	381
location_class	0	216939
road_surface_cond	0	326
road_class	0	34711
traffic_control	0	1564
visibility	0	3693

Table 2. Number of missing and empty information in the involved persons data

variable	sum_missing	sum_empty
collision_no	0	0
rec_id	0	0
accdate	0	NA
actual_speed	14830	NA
driver_condition	0	15870
impact_location	0	4192
involved_injury_class	0	6
involved_age	384	NA
involved_class	0	0
manoeuver	0	15506
pedestrian_action	0	1126
pedestrian_condition	0	169
pedestrian_collision_type	0	1455
posted_speed	14110	NA
vehicle_class	0	12372

For the involved_persons data, the columns, actual_speed, driver_condition, posted_speed, impact_location, manoeuver, and vehicle_class, have a significant missing or empty information problem. Similarly, the columns, district, location_class, and road_class, in the collisions data have significant missing or empty information.

B. Data Cleaning

I generated a preprocess function (in the function script, R/functions.R) which is used to clean the data. First, I dropped all the columns suggested in the data discovery step above. I dropped the remaining columns, which had some amount of missing or empty information. Subsequently, I checked for duplicate rows in both data. There were no duplicate rows in the collision data. However, there were significant duplicates in the involved_persons data. For each duplicate(s) in the involved_persons data; I retained only one row. Also, in this step, I created the outcome variable (serious_fatal) from the involved_persons data (specifically, the involved_injury_class column). The serious_fatal equals 1 if the involved_injury_class is fatal or major; it was converted to a factor variable.

C. Data Transformation

First, I performed a left join operation on the two data sets, using the collision_no column as the common id. The number of rows for the new data is 1,595. Also, I created 10 binary variables from the existing columns. The variables are defined as follows:

1. age_65_plus: equals 1 if the pedestrian's age is 65 or more and 0 otherwise.
2. impaired_pedestrian_condition: equals 1 if the pedestrian has been drinking or has impaired ability from drugs or alcohol or has been drinking or has ability impaired with alcohol over .80, and 0 otherwise.
3. pedestrian_inattentive: equals 1 if the pedestrian condition is inattentive.
4. collision_midblock: equals 1 if pedestrian_collision_type is pedestrian hit at mid-block.
5. collision_inter_wrow: equals 1 if pedestrian_collision_type is the vehicle going through the intersection while the pedestrian crosses without a row or vehicle turns left while the pedestrian crosses without row at the intersection or the vehicle turns right while the pedestrian crosses without row at the intersection.

6. collision_inter_row: equals 1 if pedestrian_collision_type is the vehicle going through the intersection while the pedestrian cross with row or the vehicle turns left while the pedestrian crosses with row at the intersection or the vehicle turns right while the pedestrian crosses with row at the intersection.
7. collision_transit: equals 1 if pedestrian_collision_type is a pedestrian involved in a collision with a transit vehicle anywhere along the roadway.
8. weather_bad: equals 1 if visibility is "fog, mist, smoke or dust" or strong wind or drifting snow or freezing rain.
9. no_traffic_control: equals 1 if traffic control is no control.
10. dark: equals 1 if light is dark artificial or dark or dusk or dusk artificial or dawn artificial or dawn.

In addition, I dropped the other variables except for the outcome variable (serious_fatal), latitude, and longitude. These variables will be used to estimate the probability that an accident will be fatal or major. Again, a correlation heat map showed that might not be a multicollinearity problem in the model estimation phase. A Pearson's correlation map reveals that multicollinearity might be a problem in model estimation. To generate robust performance metrics, I generated both a test set and a validation set. First, 20% of the data was set aside as a test set and the remaining 80% as a train set. Further, 20% of the train data was set aside as a validation set; the model was trained on the remaining train data.

D. Data Validation

All the selected variables for model estimation have expected values, accurate and consistent. The problem is that the level (pedestrians with major or fatal collisions) is underrepresented in the outcome variable (11% of the total sample, Table 2). Model estimation with this data will produce bias and unreliable inference. To solve this problem, I employed a SMOTE algorithm to re-balance the data. After the data re-balancing, the data set produced a relatively balanced outcome distribution (pedestrians collisions that resulted in major injury or fatality accounted for 24% of the rows of the dataset.

Summary The major data quality issues were missing or empty information and imbalanced data regarding the classes in the outcome variable. For this reason, in particular, the imbalance data problem, I do not expect the model to provide better performance metric statistics and statistical inference. However, the study could provide important qualitative insights to help in enhancing road safety in the City of Toronto.