Course: GDAA 1001 - Fundamentals of Geospatial Data Analytics

Determining Risk Factors for Injuries and Fatalities In Road Collisions

Keegan MacDonald, Sanita Richards, Corey Thorne, Reuben Walker

December 5th, 2023

**Introduction**

Road collisions remain a significant concern globally, impacting lives and posing a considerable societal challenge. Canada, with its vast network of roadways, experiences its share of incidents, contributing to injuries and fatalities. Understanding the factors associated with these collisions is crucial in formulating effective preventive measures and enhancing road safety protocols.

This report aims to delve into Canada's national collision data from the year 2019 to discern factors which may predict the severity of injuries and fatalities in road collisions. By employing data analytics techniques and modeling in R Studio, this study endeavors to uncover patterns, correlations, and potential causal relationships among various variables involved in these incidents.

**Data Preparation**

The data utilized was retrieved from the National Collision Database (NCDB). This dataset encompasses a range of categorical variables, including the month, weekday, and hour of collisions, the number of vehicles involved, road type (intersection, passing lane, bridge, etc.), and weather conditions, among 18 others. It's important to note that many variables include extra categories accounting for unknown or missing data, which required careful handling.

Our dataset was initially comprised of 272,301 observations of 23 variables. To ensure the integrity of our analysis, we judiciously removed all observations containing unknown or missing values. Subsequently, six columns deemed irrelevant for machine learning were excluded, including the redundant 'year 2019,' three columns containing unique collision identifiers, a duplicate of the target variable denoting collision severity, and a column detailing protective clothing and safety devices, largely populated with missing values. This meticulous curation resulted in a refined dataset, now comprising 17 columns and 167,803 observations, poised for robust machine learning model development and analysis.

The modified dataset underwent a meticulous cleaning and preparation phase to ensure its suitability for analysis. Categorical variables were appropriately grouped and transformed for streamlined analysis in R Studio.

The 'numCars' column was created by categorizing the 'C\_CONF' variable based on collision configurations, distinguishing between single-vehicle incidents and various two-vehicle scenarios. Conversion of several columns into factor types was conducted to enable categorical analysis, including 'P\_ISEV' (Injury Severity), 'C\_MNTH' (Month), 'C\_WDAY' (Weekday), and others deemed vital for categorical assessment.

To enhance readability and interpretability, categorical codes were reassigned meaningful labels. For instance, numerical codes representing injury severity, months, weekdays, and weather conditions were relabeled into descriptive categories. Additionally, column names were updated to ensure clarity and consistency throughout the dataset.

Subsequently, the categorical variables were grouped, optimizing the dataset for analysis and model development. Moreover, a subset of 5000 rows was randomly sampled from the prepared dataset to facilitate manageable yet representative analyses

This rigorous data preparation process resulted in a refined dataset poised for in-depth analysis, machine learning modeling, and exploration of risk factors associated with injuries and fatalities in road collisions within Canada's 2019 national collision data.

tidying (i.e., pivoting)

trimming outliers (if necessary)

normalizing or standardizing data (if necessary)

dataset subsetting (rows or columns)

handling of missing data (imputation or subsetting)

deriving new variables (e.g., making count-based data relative, if applicable)

Exploratory Data Analysis

Descriptive statistics

Examination of variation

Covariation

Predictive Modelling

Evaluation of Results

Conclusion

References (if applicable)