CIND 820: Capstone Project

Analyzing and Predicting Traffic Collisions in Ottawa

Student: Pak Him LEUNG AH KANG

Student ID: 500866890

Supervisor: Tamer Abdou

Date: Oct 28, 2024

Ryerson University

Table of Contents

Abstract

There are around 1500 to 2000 traffic injuries, 20 to 30 fatalities and 13,000 reported vehicular collisions alone in the city of Ottawa during the year 2021[1]. According to the nationwide Canadian Motor Vehicle Traffic Collision Statistics: 2021 report, the number of fatalities derived from traffic collisions rose by 1.3% from 2020 and reached a record 1786 fatalities at that time[2].

The goal of this project is to perform data analysis on the several factors that contribute to the occurrence of collisions and fatalities such as speed, weather, road surface condition, etc and using multiple models such as logistic regression, ensemble methods and decision trees to perform the analysis.

The second goal is to create a online web application accessible to the public that maps out the most high risk collision areas in the city of Ottawa and which also predicts the collision probability at a given intersection or with a user input travel route into low, medium and high risk using algorithms such as Random Forest, decision trees and support vector machines.

The datasource will come from the traffic collision data for the year 2017-2022 hosted by Ottawa's public data source (open data ottawa) and various APIs like weather and fuel costs. Next, the collected data will be combined and cleaned and an exploratory data analysis will be performed on it to determine the patterns and relationships between the combined attributes and formulating hypotheses for regression or classification techniques.
Some of the research questions include what is the correlation between the number of collisions and the external factors such as road condition, weather, time, etc.. Another research question is the examination of traffic control systems to check if they affect the number of collisions. The last question is whether the number of collisions and the severity of the collisions can be predicted at a specific intersection.

The source code will be hosted on https://github.com/mkleung/traffic-collision-analysis

Introduction

Traffic collision analysis has long been a tool used by city engineers to improve road safety and reduce the number of accidents on the roads. The recent advent of data science and artificial intelligence has enabled them to apply machine learning techniques to predict the number of traffic collisions at specific areas of the city and offer important insights and advice for urban planners and city governments as well as members of the public.

The main motivation for this project is to use data science to help reduce the costs associated with vehicular traffic collisions which include healthcare expenses, property data, lost productivity and most importantly the loss of life and the sense of security while driving. As cities in Canada are struggling with infrastructure costs, reducing traffic accidents would be beneficial in saving funds that can be redirected to other purposes.

Stakeholders

There are many stakeholders involved in traffic control. Urban planners design roads in order to make them collision free as well as the police department who will need the data analysis in order to direct funds into the proper departments. For example, in high risk collision areas, more police should be deployed in that area and at different times of the day. The data analysis will be invaluable to government officials both provincial and federal in the analysis of trends of traffic collisions over time. Additionally, car insurance companies can use the data to assess risk, predict insurance rates and fraud reduction. Also members of the public will benefit from this data, especially those who enjoy defensive driving.

Research Questions

There are six main research categories that have been chosen in this data science project. They explore various factors influencing traffic collisions in the city of Ottawa. The categories are environment, speed, traffic control, fuel costs, location and accident type. Under each category, we seek to identify correlations that can inform us about the relationship between the category and the number of traffic collisions.

1.	**Environment** How do environmental factors such as weather, snow, rain and fog affect the road surface conditions that lead to the frequency of collisions?
2.	**Speed** Does vehicle speed influence traffic collisions?
3	**Traffic Control** Do traffic control measures such as red light cameras, speed cameras, speed display radar and speed bumps reduce collisions?
4	**Fuel Costs** Do fuel costs for gasoline and diesel affect traffic collisions?
5	**Location** Which locations and types of roads such as highways have the highest number of fatal and non fatal accidents in Ottawa?
6	**Accident Type** Which types of accidents like rear-end or side swipe are the most common? Are they related to the location and the type of road such as highways and less busy streets?

Literature Review

Paper 1 - Effects on inclement weather events on road surface conditions and traffic safety: An event based empirical analysis framework
Abohassan et Al. 2022

Abohassan et Al. conducted a study on the relationship between pavement friction coefficient and collision counts during inclement weather in Edmonton, Alberta during the years 2017-3029 and it shows how snowy road surface conditions in -15 degrees celsius for extended periods of time in late January increase the risk of collisions by 1,091% in injury crash rates and 2,113% increase in non-injury crash rates. The researchers used hourly weather datasets along with negative binomial safety performance functions to create models and found a statistically significant relationship between road friction and traffic collisions.

The study puts forward the idea that the environmental variables such as snow and ice decreases road friction to below 0.35, leading to a rise in traffic collisions. It also mentions larger roads such as main roads, motorways and highways with higher traffic volumes and speed limits contribute to higher traffic collisions compared to lower traffic smaller roads and neighbourhood streets. The paper concludes that a strong statistically significant relationship exists between pavement friction and collision counts and that if a severe snowstorm is forecasted, preventive measures have to be broadcasted to the general public.

Even though the paper used snow as its main focus, other extreme conditions such as low visibility situations such as early morning fog, lightning and heavy rainfall directly affect the risk of traffic collisions. The study is directly related to the research question which asks if weather has any effect on traffic collisions.

Paper 2 - Examining the relationship between wind speed, climatic conditions, and road accidents in Iran

Mahshid Eltemasi et al. 2024

In Iran, a study by Eltemasi et al. (2024) looked into the relationships between the different causes of road accidents ranging from distracted driving, wind speed, excess speed, human fatigue, environmental conditions and focusing particularly on wind speed using road accident data and wind speed data. The goal of the study was to identify which attributes or variables are the primary causes of car accidents and to provide recommendations for policy makers.

The study looked into road accident data from 2017 to 2022 consisting of 15 thousand cases and also wind speed data and used logistic regression to assess the relationship between wind speed and the type of accidents. The logistic regression results concluded that wind speed has no effect on the category of accidents like fatal or non-fatal but the only wind speed influence has is in the way the vehicles collide. They also used data mining like the J48 decision tree to visualize the relationships between numerous weather related attributes. The decision tree was able to predict collision probabilities in road accidents and found that rainy conditions increase the probability of non-fatal accidents and higher wind speed increases the probability of fatal accidents in front-to-back collisions.

A number of accident experts have been interviewed as well as past texts and articles have been reviewed which reveal human factors, not just environmental factors, contribute significantly more to traffic collisions. Fatigue and distractions such as mobile phones are the primary contributors to collisions while environmental factors come at a close second. The paper is directed related to the first research question and directly addresses whether the environmental factors contribute to traffic accidents.

Paper 3 - Reducing Road Traffic injuries: Effectiveness of speed cameras in an urban setting

Perez et al. 2007

A study investigated the effectiveness of speed cameras installed in Barcelona (Perez et al. 2007) at reducing traffic collisions between two groups using a local police database. The first group was conducted on the beltway of Barcelona which is a major highway without any speed cameras and the second group was conducted on the same beltway with new traffic cameras installed. The goal of the study was to find out if the installation of speed cameras led to a reduction in collisions and injuries.

The experiment conducted was based on a time series design and they used Poisson regression models to deal with historical trends and seasons to find that there is a 27% reduction in road collisions since the traffic camera installation. There was a reduction in injuries by 507 and 789 fewer vehicles were involved in these collisions. The study also found that the cameras prevented collisions at all times of the day, including nighttime and also on weekends.

This paper is directly connected to the research question number 3 which asks whether traffic cameras have an effect on collisions. It focuses on the effectiveness of speed cameras and provides empirical evidence on the impact of reducing road collisions, especially using two groups, one without and one with speed cameras installed after. Although the study’s main focus is on speed cameras, other traffic control measures such as red light cameras, speed bumps, speed display radars and even the presence of police cars will have an effect on reducing traffic collisions.

Paper 4 - Safer Roads Owing to Higher Gasoline Prices: How long it takes
Guangqing Chi et al. 2015

This paper by Guangqing Chi et al. studies the correlation between gasoline prices and traffic collisions with a special consideration on the time between a change in gasoline prices and its effect on traffic collisions using the 2004 to 2023 traffic crash data and several variables such as economic difficulty, seat belts, alcohol consumption and fuel costs data from the US Department of Energy Information Administration.

The goal of the study consisted of three goals which were to find the positive relationship between higher gasoline prices and traffic collisions. The second goal was to check if variables like age, gender and race correlates with traffic accidents and the final aim is to identify how many months passed before gasoline prices have any effect on collisions which will be our main focus because it is related to research questions number 4.

The study used negative binomial regression and Pearson correlations to investigate how fuel costs affect traffic crashes. The researchers found out that the effect of gasoline changes on traffic collisions did not occur until 9-10 months after using a negative binomial regression model. The researchers also found that gasoline prices have different effects on different age groups. For teenage drivers, gas prices have an immediate effect on traffic collision reduction and this age range is the only group that has been affected by gas prices. There are no effects on adult drivers.

The findings concluded that higher gasoline prices lead to a smaller number of traffic collisions and suggested that government officials and policy makers should consider increasing gasoline prices in order to reduce traffic collisions.

Paper 5 - Factors influencing accident severity: An analysis by road accident type
Laura Eboli et al. 2007

This paper by Laura Eboli et al. conducted research on the different types of traffic collisions such as front/side and rear end collision and the relationships between the type of collision and the number and seriousness of injuries. They analyzed a dataset of 40, 172 road accidents that took place in Italy in 2016 and focused on three characteristics that might contribute to traffic collisions such as the type of road, weather conditions and the driver details such as his or her age, gender and license type.

Using a binary logistic regression, they modeled the relationship between accident type and various categorical attributes and tried to determine the importance of each attribute's effect on traffic collisions. The findings showed that there are significant differences in accident severity and the type of collisions. The location of the collision is a major factor which include intersection type. Other factors include driver related relationships such as license type, experience directly influence the severity of accidents. Gender has no impact on accident severity.

This study is interesting because it directly answers the last two research questions of this capstone project.

Data Description

This project’s dataset is available through the open Ottawa's government website which is a web portal that offers datasets related to the city of Ottawa like details about traffic, environmental and housing data amongst others. It is called “Traffic Collision data” and is one big csv file and contains data about road surface, weather, type of collisions, number of injuries, etc for the year 2017-2022 and the dataset can be found in the references.

The dataset contains 74,612 total number of records and 30 Columns or attributes. Here is a list of attributes at first glance and their corresponding descriptions.

- Date
- Time
- Location (RD1 @ RD2 or RD from RD 1 to RD 2)
- Location Type (Intersection, non-intersection, at/near private driveway)
- Classification of collision (non-fatal, fatal, property damage only)
- Initial impact type (Angle, turning movement, rear-end…)
- Road surface condition (Ice, wet, dry snow...)
- Environment (Clear, rain, snow…)
- Light (daylight, dawn, dusk…)
- Traffic control (stop, traffic signal, no control…)
- Number of Vehicles
- Number of Pedestrians
- Number of Bicycles
- Number of Motorcycles
- Max Injury (Highest injury level in the collisions)
- Number of Injuries
- Number of Minimal Injuries (Person did not go to hospital when leaving the scene of the collision)

- Number of Minor Injuries (Person went to hospital and was treated in the emergency room, but not admitted)
- Number of Major Injuries (Person admitted to hospital. Includes person admitted for observation. This could be either life threatening or non-life threatening)
- Number of Fatal Injuries (Person killed immediately or within 30 days of the motor vehicle collision)
- X and Y Coordinate (MTM Zone 9, NAD83)
- Latitude and longitude (WGS1984)

Here are some statistics about the data. There are 17 numeric and 13 categorical columns. Running a null checker scan using python reveals the following details. Accident_year consists 100% of missing values. A few missing null values (less than 5) are in Accident_Year, Initial_Impact_type, Road_Surface_Condition, Traffic_Control and Environment_Condition.

```
Descriptive Statistics for Numeric Columns:

                X             Y  Accident_Year  Num_of_Vehicle  \
count  7.461200e+04  7.461200e+04            0.0    74612.000000
mean  -8.428035e+06  5.670130e+06            NaN        1.841219
std    1.870145e+04  2.316663e+05            NaN        0.586512
min   -8.820655e+06  0.000000e+00            NaN        1.000000
25%   -8.433028e+06  5.674170e+06            NaN        2.000000
50%   -8.426515e+06  5.681383e+06            NaN        2.000000
75%   -8.420483e+06  5.687619e+06            NaN        2.000000
max   -8.378081e+06  5.704542e+06            NaN       25.000000

       Num_Of_Pedestrians  Num_of_Bicycles  Num_of_Motorcycles  \
count        74612.000000      1347.000000          637.000000
mean             0.022289         1.010393            1.015699
std              0.153846         0.115173            0.136459
min              0.000000         1.000000            1.000000
```

```
25%              0.000000         1.000000            1.000000
50%              0.000000         1.000000            1.000000
75%              0.000000         1.000000            1.000000
max              3.000000         3.000000            3.000000

       Num_of_Injuries  Num_of_Minimal_Injuries  Num_of_Minor_Injuries  \
count     13417.000000              5733.000000            7804.000000
mean          1.298353                 1.212977               1.226422
std           0.734951                 0.562083               0.585715
min           1.000000                 1.000000               1.000000
25%           1.000000                 1.000000               1.000000
50%           1.000000                 1.000000               1.000000
75%           1.000000                 1.000000               1.000000
max          38.000000                11.000000              10.000000

       Num_of_Major_Injuries  Num_of_Fatal_Injuries   X_Coordinate  \
count             671.000000             141.000000   74612.000000
mean                1.108793               1.070922  366572.012158
std                 0.612139               0.283987   15213.451923
min                 1.000000               1.000000       0.000000
25%                 1.000000               1.000000  363147.689000
50%                 1.000000               1.000000  367676.548050
75%                 1.000000               1.000000  371925.281000
max                14.000000               3.000000  401821.931000

       Y_Coordinate           Lat          Long      ObjectId
count  7.461200e+04  74612.000000  74612.000000  74612.000000
mean   5.017227e+06     45.291930    -75.710325  37306.500000
std    2.042363e+05      1.843572      0.167998  21538.773479
min    0.000000e+00      0.000000    -79.237290      1.000000
25%    5.021766e+06     45.333451    -75.755176  18653.750000
50%    5.026853e+06     45.378984    -75.696671  37306.500000
```

```
75%    5.031266e+06     45.418314    -75.642484  55959.250000
max    5.043439e+06     45.524921    -75.261583  74612.000000
```

Outliers

There are some potential outliers

Y (Coordinate): Minimum value of 0

Num_of_Vehicle: Maximum value of 25

Num_Of_Pedestrians: Maximum value of 3

Num_of_Injuries: Maximum value of 38

Num_of_Minimal_Injuries: Maximum value of 11

Num_of_Minor_Injuries: Maximum value of 10

Num_of_Major_Injuries: Maximum value of 14

Num_of_Fatal_Injuries: Maximum value of 3

X_Coordinate and Y_Coordinate: Minimum values of 0

Data Cleaning

Firstly, the column Accident_Year which does not contain any data and then the attributes that also need to be removed are the ID, X, Y, ObjectID, X_Coordinate and Y_Coordinate. These attributes are possibly used by internal systems and therefore deemed irrelevant to our study.

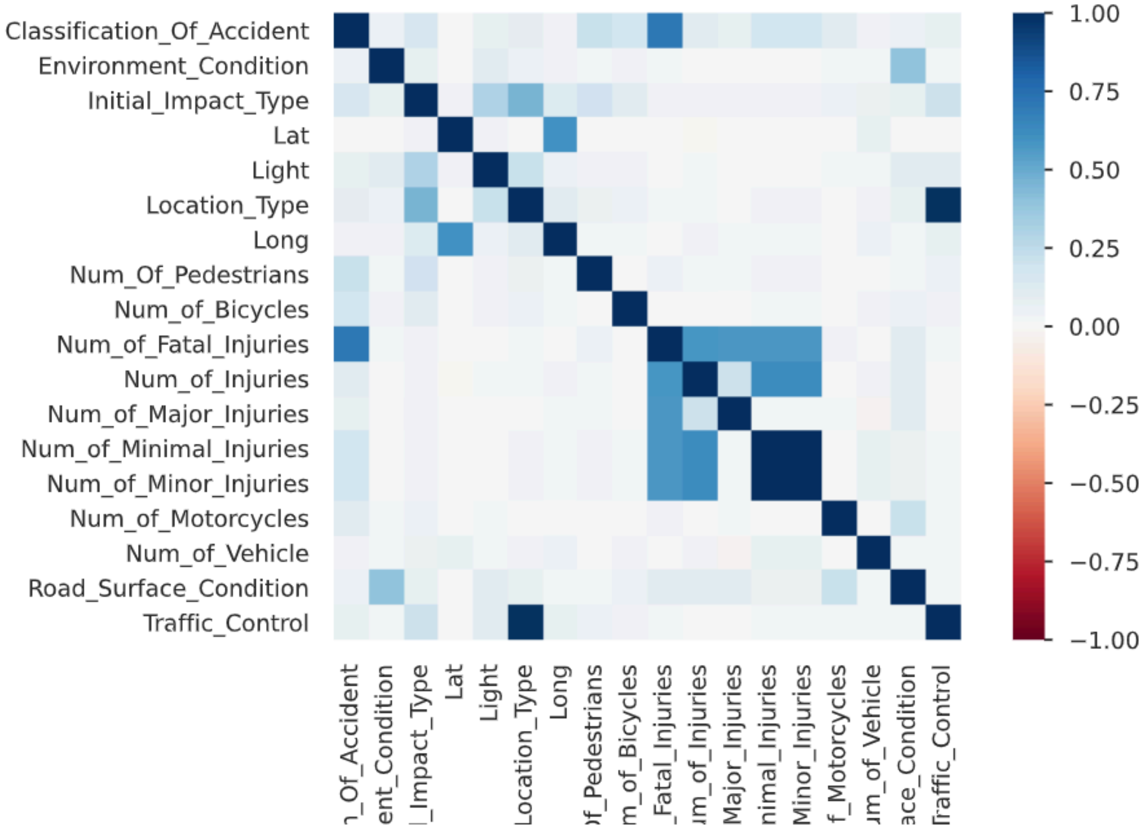
Next we need to combine the Accident_Date and Accident_Time into a single column as having two separate columns for date and time is redundant.

We will then fill a large number of missing values in the attributes like Num_Of_Bicycles, Max_Injury and Num_Of_Motorcycles with zero because it is assumed that if there is no value, it means there are no number of bicycles or motorcycles involved in the accident and Ottawa is mainly a car centric city.

There are also a number of minor missing values (less than 5) in several attributes such as Initial_Impact_type, Road_Surface_Condition, Environment_Condition and Traffic_Control which can be replaced by a mode.

The following data frame has been generated with the following 19 attributes after cleaning

Attribute	**Type**	**Attribute**	**Type**
Accident_DateTime	datetime	Num_Of_Motorcycles	discrete
Classification_Of_Accident	nominal	Max_Injury	discrete
Initial_Impact_Type	nominal	Num_Of_Minimal_Injuries	discrete
Road_Surface_Condition	nominal	Num_Of_Injuries	discrete
Environment_Condition	nominal	Num_Of_Minor_Injuries	discrete
Light	nominal	Num_Of_Major_Injuries	discrete
Traffic_Control	nominal	Num_Of_Fatal_Injuries	discrete
Num_Of_Vehicle	discrete	Lat	continuous
Num_Of_Pedestrians	discrete	Long	continuous
Num_Of_Bicycles	discrete		



The correlation matrix

Here is a summary of the correlation matrix

- There is a strong correlation between number of injuries types with each other, meaning one type of injury tend to have a greater chance of having multiple types of injuries
- There is a high correlation between number of fatal injuries and number of other injuries, meaning for one fatal injury tend to have multiple other kinds of injuries as a result
- Traffic control and environment condition have weak correlation with injury counts

Abnormal data

The data inside the location field consists of two road names separated by the symbol @ or occasionally by the word “btw”. For example, WEST RIDGE DR btwn PARLOR PL & BERT G. ARGUE DR (__5RG32N). This field should be removed as this is unnecessary.

Approach

The first goal of this data science project is to understand the relationship between different attributes such as weather, time, etc with the number of collisions. A good approach will be to use decision trees and random forest classifiers to visualize the relationships between the different attributes and collision.

The second goal will be to conduct some predictive analysis on the severity or type of accidents based on attributes such as traffic control in order to answer if traffic control systems affect collisions in Ottawa. Again, decision trees and random forests would be the best methods to conduct the analysis.

The final goal of this project will be to use geospatial analysis to identify the locations where accidents are the most prevalent. Clustering techniques such as K-means clustering will be used.

References

Dataset Source https://open.ottawa.ca/datasets/ottawa::traffic-collision-data/about

[1] Ahmed Abohassan et. al. retrieved from
https://journals.sagepub.com/doi/full/10.1177/03611981221088588

[2] Eltemasi et al. (2024) retrieved from

https://www.cell.com/heliyon/fulltext/S2405-8440(24)09259-4

[3] Perez et al. 2007 retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC1963295/

[4] Guangqing Chi et al. retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC4504271/

[5] Laura Eboli et al. retrieved from

https://www.sciencedirect.com/science/article/pii/S2352146520303197