Data Science Capstone Project

Table of Contents

Introduction

Background

From the 2019 Seattle Car Accident Statistics & Reports: The Washington State Department of Transportation (WSDOT) reports that in 2019, there were 10,315 total collisions in Seattle; of those, 3,658 resulted in possible injuries. They go on to report that deaths are up but serious injuries are down (part of this relating to the Ride the Ducks vehicle crash in 2015 which resulted from improper maintenance.

Problem:

Seattle's collision rate from 2015 to present has been on the rise and with more collisions comes the likelihood of more accidents.

Project Intent:

The intent of this project is to attempt to identify situations which are more likely to result in injury.

Data Description/Analysis

Summary

After analyzing the data, it appears that the collision type and weather are the most significant fields in terms of correlation to the severity of an accident. However, the weather condition is more likely to be incidental and the collision type appears to be the main cause of the severity being that nine out of ten times a collision with a pedestrian or cyclist results in injury.

Raw Data:

- The given csv file contained 38 fields, of these, thirteen were integers, four were floating point/decimal numeric values, and the rest were mixed data or text.
- Upon analysis of the raw data it became apparent that some cleanup needed to be done. NaN values (missing data/null) were replaced with zeros, and fields which may have had zeros and "N"s within

the same field or 1s and "Y"s in the same field needed to be standardized. Values within these fields were all replaced with either 0s or 1s to be used as ints or bools.
- Missing values (NaNs) in text fields were replaced with the text "Unknown."
- Two fields (X, and Y) appear to be latitude and longitude. These may be useful in determining if certain areas of the city are more prone to accidents.

Analysis:

The data was grouped into two sets, numeric and text, to determine where there was significant variation between the values within the fields.

Numeric data fields
('SEVERITYCODE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','UNDERINFL','SPEEDING','HITPARKEDCAR'):

- **Severity** - Only two values exist within the example dataset but the metadata document indicates that there are five possible values for severity (0-Unknown,1-prop damage,2-injury,2b-serious injury,3-fatality). For further analysis of data including the remaining codes, the data would have to be converted to strictly integer codes.
- **Person Count/Vehicle Count** - appears to have no significance on the severity of the accident. The vast majority of collisions include only a single person and therefore are highest for a single person for both property damage and injury.
- **Ped Count and Ped Cycle Count** - Severity appears to be higher (above average and more likely to result in injury) as pedestrians and cyclists have less protection when struck by a vehicle.
- **Under the Influence and Speeding** - May result in a severity higher than two but unable to verify without data which contains severity codes greater than two.
- **Hit Parked Car** - Overwhelmingly property damage; these incidents are likely to occur at low speeds (while parking) and are unlikely to result in injury.

Text data fields
('ROADCOND','COLLISIONTYPE','WEATHER','LIGHTCOND'):

- **Note:** The text data fields seem to offer the best correlation between severity and their individual values; particularly for weather conditions and collision type.
- **Road Condition** - low correlation between the road conditions and the severity of the accident. Based on the data, one may be able to say that there is a slightly higher likelihood that wet or oily road conditions have slightly higher effect on the severity of an accident but overall, there are more significant correlations based on Collision Type and Weather.
- **Collision Type** - Collisions with cycles and pedestrians have a much higher probability of being an injury (sev type 2). Almost nine out of ten collisions with pedestrians or cyclists result in injury.
- **Weather** - Oddly enough, the weather condition 'Partly Cloudy' in most incidents where severity indicates injury. While there is a correlation, this does not imply causation and is most likely due to the high number of partially cloudy days in Seattle.
- **Light Condition** - Does not seem to be a significant factor in determining collision severity. Even when replacing 'Dusk' and 'Dawn' with 'twilight' there's no significant change in average severity of collisions.

Methodology

Raw Data Investigation/Analysis

I initially started with data exploration. It's important to understand what information (fields) is available in the dataset as well as the values that can be found in each field. For that I checked the datatypes for each field, viewed first five records of each type and started to determine how these would be evaluated. I decided not to focus on the longitude and latitude at this time due to time constraints and not knowing Seattle well enough to see what may be causing changes in severity due to location. After reviewing the data, I selected 15 fields that I thought would best suit my analysis and provide the clearest picture of the factors relating to the severity of a collision.

Data Hygiene/Cleanup and Preprocessing

Any given raw set of data that's input by humans will have a significant amount of variation within any given field. Spelling errors, different understanding of what's expected for input, etc. So, before beginning the process of analyzing the data it's important to do some cleanup. Data within fields should be standardized for text capitalization, numeric data should be standard and without text values, and fields that should be true or false should be standardized so that they do not include a variety of methods that people use to indicate yes/no, true/false, x/null, 1,0, etc. For this dataset I replaced all NaNs (nulls) with zeros for numeric fields and 'Unknown' for all text fields. All 'N's or 'Y's (or other variations for Boolean fields) were replaced with integers being either 0 or 1 to represent false or true respectively.

Histograms, Severity Means, and Grouping

For the strictly numeric data I decided to view the values as histograms. It was a simple process to display histograms for each of the numeric values and quickly see which may have significant variation that may need to be explored. No significance was found within the numeric values.

The four text-based fields provided more relevant information but had to be processed differently. For these I explored the means of severity for each along with the normalized groupings of severity by the values within each text field. For instance, the average (mean) of severity for partly cloudy was 1.6000. As this value is above 1.5 (the median between severity 1 and 2), this indicates that injury is above average on partially cloudy days. Upon exploring the normalized grouped-by counts of severity for partially cloudy we see that 60% of the collisions on partially cloudy days result in injury. The same analysis was done for all text-based fields selected.

KNN tests for Severity relations to Weather and Collision Type

Upon determining that collision type and weather were potentially significant factors in the severity of incidents I decided to attempt to classify them with K-nearest neighbor. Running through all the k values for weather and collision type it was determined that k of 8 gave the best results for weather and k of 4 provided the best results for collision type. Overall, the accuracy of KNN was about 75% for both weather and collision type.

Results

After analyzing the numeric fields as well as the quantitative fields (text) it looks like the most likely factor in determining the severity of an accident is the collision type. Vehicle collisions involving pedestrians or cyclists result in injury about 90% of the time. Given the example dataset it's hard to determine what other factors would have a significant effect on collision severity.

Discussion

During the analysis of the data I starting acquiring additional questions. See below. I was only able to answer a few of them because some of the questions would require additional data or perhaps because some of the points of data had so few instances that they would tend towards the law of small numbers (smaller samples have greater variability and therefore less accuracy).

Questions

- Are there incidents of collisions with pedestrians or cyclists where there is no injury?
- What is the rate of collisions under the influence where there is no injury?
- What happened in the instances where there were more than 10 people in an accident? (outliers?)
- Number of incidents where speeding is involved is low, but what's the rate of injury where it does happen?
- Number of incidents where a pedestrian is involved is low, but what's the rate of injury where it does happen?
- What about collisions where no vehicle is involved?

Overall i think only the questions involving pedestrians and cyclists were answered and given their resulting injury rate I think it's fair to say that if you just assumed that any time a vehicle collides with a pedestrian or cyclist, you'd have better accuracy than was achieved using KNN.

If I was to re-evaluate this data I might start by removing the pedestrian and cycle counts to see what the next most important factors would be in determining the severity of accidents.
What about the instances of collisions while speeding or where people are under the influence?
I think we'd just have to have more data to determine that. given the size of the data the instances of speeding and under the influence are very low. Additionally, I think we'd need better data around the severity of those collisions. How many were fatalities? How many are classified as serious injury?

Conclusion

Overall, we need to define not just one, but a whole host of questions that come together to give us a more well-rounded perspective on the nature of vehicle collisions and the factors that attribute to the severity of their impacts.

From the given data we can say that the largest factor in severity has to do with what the vehicle impacts and whether or not pedestrians or cycles (individuals who are not protected by being inside vehicles) but we do not have enough information to define a better classification system for a sliding scale of severity. In the data given the reality is that there were only two options for injury; injured or not injured. Given what we know about this data I think the case is simply that people are fragile.

Backup Information (Python/Jupyter Notebook)

Below you will find the Python code that was used to do the analysis.

View Notebook here -> https://github.com/Zadigrim/DSPC_Capstone