Predicting the Location of Severe Accidents

John Papalia

October 9, 2020

# Introduction

## Background

Accidents happen. While accidents, by definition, do not happen purposefully, it is possible to look at existing or potential situation and evaluate whether that situation is more or less likely to be a breeding ground for accidents. By using an analysis such as this, city planners may be able to implement factors to mitigate the occurrence or severity of these accidents.

## Problem

Data that can contribute to this analysis includes accident reports and accident information, accumulated over a historical period. By looking at the features of the accident location, the types of accidents (car vs human? car vs pedestrian?), and any other available factors, it may be possible to create a predictive model, pointing at areas of a city that should be investigated further.

## Why?

Accidents can be destructive. Destructive to property, but more so destructive to lives and families. If any large scale or incremental steps can be made to reduce accidents, then by extension the destruction to lives and property can also be reduced.

# Data Acquisition and Cleaning

## Data Sources

The information provided with the course serves as the data set for this report. The data is provided by the city of Seattle and represents a small subset of their full data set. From the onset, there is a noticeable flaw in this data set. The “Severity Code” field can be a value of 1, 2, 2b, or 3 with 1 being the least severe and 3 being the most severe. This sample data set consists of Severity Codes equal to only 1 and 2. There are no values of 2b and 3. Given that the supplied data set spans over 14 years, this omission appears glaring (or, perhaps, was purposeful?). As such, my first step was to search for, and obtain, the full data set from the city, located [here](https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions/data). The starting data set contained 221,739 records, dating back to the early 2000’s. This is good as it allows for spanning many seasons and takes into account varying weather patterns across two decades. This allows for good rigor in the data points that include weather related factor, such as road condition, weather condition, etc.

## Data cleaning

After the data was downloaded it was thoroughly examined in order to plan for cleaning. There were numerous issues with the data set, but none were insurmountable. First, several columns were deleted. The columns deleted, and the reason for deletion, are listed below in Table 1.

Removing these ancillary columns helped to tame the data file significantly. Further exploration of the data showed that ‘PEDROWNOTGRNT’ (Pedestrian not given right of way), ‘SPEEDING’ (speeding), and ‘INATTENTIONIND’ (driver not paying attention) were sparsely populated. They contained either a “Y” or a blank value. It was initially inferred that “if it’s blank, it must be an N”, however looking at other columns in the data, such as “UNDERINFL” (Under the Influence) and “HITPARKEDCAR” (Hit Parked Car) contained both positive and negative responses. Since this information calls the inference into question, it is instead necessary to discard PEDROWNOTGRNT, INATTENTIONIND and SPEEDING since the data set is incomplete and non-imputable.

Table 1 Summary and description of data file columns that were deleted.

|  |  |
| --- | --- |
| OBJECTID | Record identifier, not needed |
| INCKEY | Record identifier, not needed |
| STATUS | Record status, not needed |
| LOCATION | Text version of location, not needed |
| SEVERITYDESC | Text version of location, not needed |
| COLDETKEY | Another unique identifier, not needed |
| REPORTNO | Another unique identifier, not needed |
| INCDATE | Date information that is duplicated in INCDTTM, kept INCDTTM which has full time stamp |
| SDOT\_COLDESC | Each incident gets a numeric code by Seattle DOT, this is the text descriptor of the code; keep the numeric, get rid of this text version |
| SDOTCOLNUM | Another identifier that is not needed |
| ST\_COLDESC | Each incident gets a numeric code by the State, this is the text descriptor of the code; keep the numeric, get rid of this text version |
| SEGLANEKEY | Describes the lane segment where incident occurred, this data is mostly blank |
| CROSSWALKKEY | Describes the crosswalk where incident occurred, this data is mostly blank |
| EXCEPTRSNCODE | This column can contain a descriptor of “NEI” or “Not enough information”. When this NEI code appears, there is no correlated geographic location data. First, all rows with “NEI” in this column were deleted, then the column itself was removed. |
| EXCEPTRSNDESC | This is the text version of EXCEPTRSNCODE; column empty, removed. |

Since one of the main points for comparison and correlation in this analysis centers around the accident severity (SEVERITYCODE), that data in that column has high value. That column can have a number equal to 1, 2, 2b, and 3, where each higher value indicates a more severe accident. There was a subset of records where the SEVERITYCODE was zero, indicating the information was not supplied. It also could not be inferred. Since this data was missing, those rows where SEVERITYCODE was zero were deleted.

At this point, the remaining data set consisted of 194,396 rows. Further row deletions were made to remove the remaining data that did not have X/Y values (location), leaving 190,946 final records. Additional data manipulation was required in order to make the data more usable. This included actions listed in Table 2.

The fields ADDR TYPE, COLLISIONTYPE, WEATHER, LIGHTCOND, ROADCOND) contained text descriptors, some of which were very granular. In order to reduce the total number of labels, some label consolidation was performed as indicated in Table 3.

Table 2: Columns that had their data reformatted, along with description of the dataformatting changes.

|  |  |
| --- | --- |
| UNDERINFL  (Under the influence) | This column contains both Y/N and 0/1 values. 0’s were replaced with N’s, 1’s were replaced with Y’s. |
| HITPARKEDCAR  (Hit parked car) | This column contains both Y/N and 0/1 values. 0’s were replaced with N’s, 1’s were replaced with Y’s. |
| SEVERITYCODE  (Severity Code) | This column contained values 1, 2, 2b, 3 and was of object type. To advance further analysis, the values of 2b were converted to 2.5 (maintaining the severity hierarchy). |
| WEATHER  ROADCOND  LIGHTCOND | Contained a quantity of “UNKNOWN” entries. In most cases, the UNKNOWN values of these three fields overlapped. All rows with UNKNOWN in these fields were removed. |

Table 3: Summary of label consolidations performed on COLLISIONTYPE, LIGHTCOND, and WEATHER columns

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Original Label** | **New Label** |  | **Original Label** | **New Label** |
| **COLLISIONTYPE** | | **LIGHTCOND** | |
| Angles | Moving | Dark - No Street Lights | Dark |
| Head On | Moving | Dark - Street Lights Off | Dark |
| Left Turn | Moving | Dark - Street Lights On | Dark |
| Parked Car | Moving | Dark - Unknown Lighting | Dark |
| Rear Ended | Moving | Dawn | Dawn-Dusk |
| Right Turn | Moving | Dusk | Dawn-Dusk |
| Sideswipe | Moving |  |  |
|  | | | | |
| **WEATHER** | | | | |
| Blowing Sand/Dirt | | | Reduced Visibility | |
| Fog/Smog/Smoke | | | Reduced Visibility | |
| Raining | | | Precipitation | |
| Sleet/Hail/Freezing Rain | | | Precipitation | |
| Snowing | | | Precipitation | |
| Overcast | | | Cloudy | |
| Partly Cloudy | | | Cloudy | |

# Exploratory Data Analysis

## Fundamental Relationships

It could be hypothesized that if most accidents would (should?) occur in adverse conditions, such as darkness or ice/snow, etc. What preliminary analysis of the data indicates is that is not the case. By far the greatest number of accidents recorded occur on days where the weather is clear, and/or during daylight hours (non-Dusk/Dawn hours). One would think this would be ideal driving times for safe driving. Further one might think that accidents are more likely to occur in an intersection, where greater entropy might be at play, but again, this is contradicted by the data set as shown below in Figure 1. In fact, the only data histogram that one might consider to be “obvious” is that a majority of recorded accidents occurred with moving vehicles. The relationships between the most severe accidents (level 2.5 and 3) are hard to discern from the data set due to their relatively small number compared to Level 1 severity accidents.

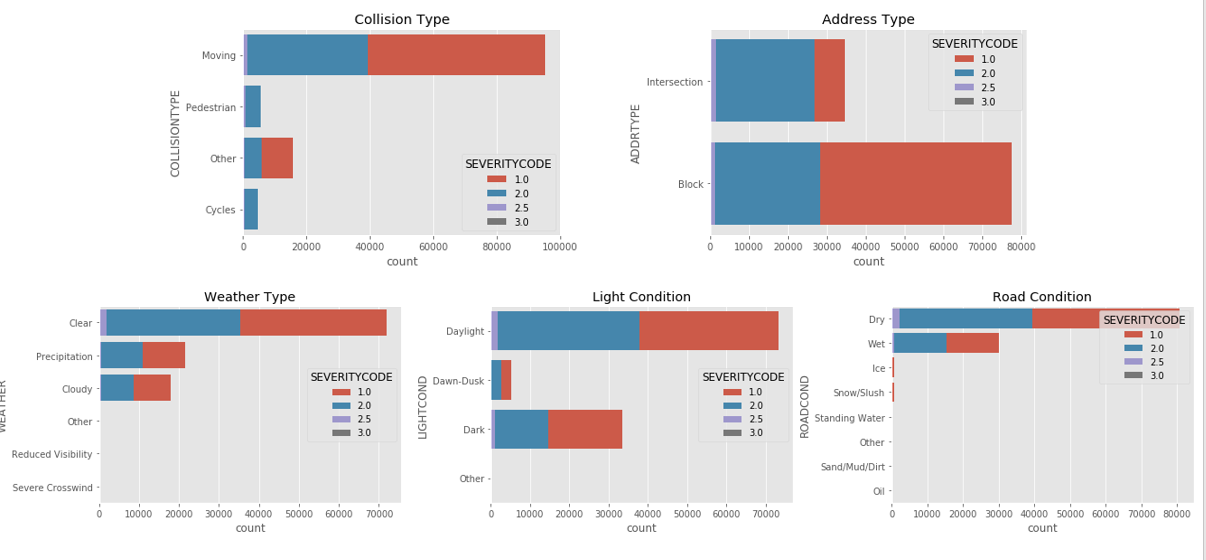


Figure 1: Stacked bar charts indicating the breakdown of number of accidents for a given severity, versus the various environmental conditions known (weather, road, etc).

## Predictive Modeling

During modelling, additional data wrangling needs were discovered. A decision tree, SVM and logistic regression model were fit to the data supplied. During model generation and testing, it was shown via confusion matrix that the severely imbalanced nature of the data set was preventing the models from fitting SeverityCode = 2.5 and SeverityCode = 3 data points. In fact, the confusion matrix yielded values of zero for both of these values. In an attempt to drive the models to better fit the full data set, the data set was balanced, reducing the number of SeverityCode = 1 data rows to equal the number of SeverityCode = 2 data rows. Unfortunately even with this data reduction, the data set is still not balanced enough to account for the higher SeverityCode values.

Table 4: Confusion Matrices for the three models (Decision Tree, SVM, Logistic Regression) used in this report.

|  |  |
| --- | --- |
|  |  |
|  | |

It is possible to compare the fit data for these three models as well (Figure 2) and what is found is that the three models fit almost identically. This is great news in terms of model consistency; however, it also goes to reinforce the idea that the model is not adequately fitting and addressing the two higher SEVERITYCODE values.

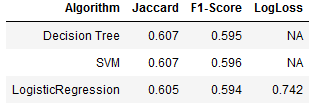


Figure 2: Summary of fit data for the models used in this work.

# Conclusions

In this report, I analyzed a date set from the City of Seattle which represents a comprehensive summary of their vehicular accident data over the past two decades. I show that the data set, while expansive, does not have an adequate enough number of higher-severity data points to be able to supply a model that is both rigorous and also able to predict the causation factors of those higher severity accidents. It could be possible to resample the data set and balance it even more than that it has been done already, but to do so will reduce the number of data points dramatically. Recreating this report with this significantly reduced data set is outside the scope of this current report (i.e., it’s another project). This follow-on project may, however, yield results that are more useful than those found in this existing report.