Capstone Project - Car accident severity

Bart Fischer

10-10-2020

## 1. Introduction & Business Problem

For a final assignment, the goal is to come up with an idea to use the provided accident data to predict the various accidents' severity. Important factors are the project validity and relevance towards the community or a group of people.

For this the **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) approach is used.

A major worldwide goal is the reduction traffic accidents, to improve road safety. Per example, in 2013 there were 1.25 million deaths around the world caused by traffic accidents. It's likely that this number will increase in the coming years.

By being able to predict accidents and the severity of the accidents before they occur one can enable safer routes, optimize public transport, and improve on infrastructure costs. This has the meets the goal to make roads and travel safer for all.

The goal of this assignment is the prediction of traffic accident severity based on a binary severity classification of accidents. The severity is classified as property damage (class = 1) or injury (class = 2). By being able to predict the outcome of the severity, one can perhaps develop saver routes (based on time/location), improve ambulance response and look into possible infrastructure problems. Thus the results are relevant for not only drivers themselves but also for public transport officials and first responders.

To summarize, the goal is to investigate a road accident dataset with a binary severity class and try to predict the severity of an accident, be it property damage or injury, by specific attributes.

## 2. Data & Data Processing

The raw data-file is 194673x38, which entails 194673 accidents with a binary severity classification, (SEVERITYCODE) and 37 additional attributes. The file represents all types of collisions from Jan-2004 to May-2020 in the Seattle (USA) area.

**What are useful attributes for modelling?**

Next to the severity classification (SEVERITYCODE - 0) there are 37 various attributes. After analysing the provided meta-data file the following selection is made with related attributes grouped.

**Relevant attributes:**

- Location related attributes: X (1), Y (2), LOCATION (10);
- Accident attributes: SEVERITYDESC (14), COLLISIONTYPE (15), PERSONCOUNT (16), PEDCYLCOUNT (17), PEDCYLCOUNT (18), VEHCOUNT (19), JUNCTIONTYPE (22), SDOT_COLDESC (24), INATTENTIONIND (25), UNDERINFL (26), WEATHER (27), ROADCOND (28), LIGHTCOND (29), PEDROWNOTGRNT (30), SPEEDING (32), ST_COLCODE (33), ST_COLDESC (34), HITPARKEDCAR (37)
- Time related attributes: INCDATE (20), INCDTTM (21)

All these attributes give some relevant information related to the accident and will be initially included to continue with.

**Non-relevant attributes:**

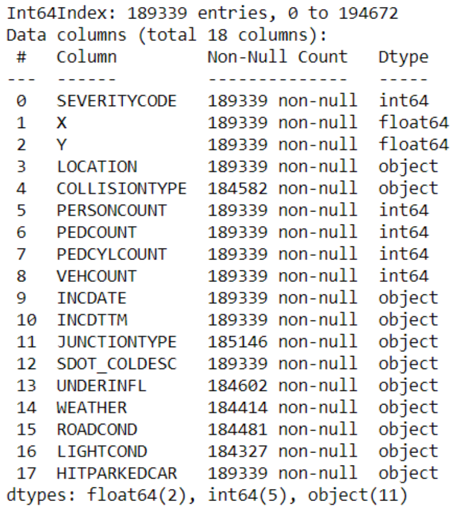
- (Unique) Keys: OBJECTID (3), INCKEY (4), COLDETDEY (5), REPORTNO (6), INTKEY (9), SDOT_COLCODE (23), SDOTCOLNUM (31), SEGLANEKEY (35), CROSSWALKKEY (36). All keys are arbitrary numbers that aren't usable for further modelling.
- None relevant data, incomplete or duplicate data: STATUS (7), ADDRTYPE (8), EXCEPTRSNCODE (11), EXCEPTRSNDESC (12), SEVERITYCODE.1 (13), SEVERITYDESC (14); All these attributes are deemed not useful to the analysis, because they are not useful to distinguish collisions, are incomplete or are a duplicate of another attribute.

All the above non-relevant attributes will be excluded from the data frame, as they hold no useful information or hold incomplete information.

As many of the remaining attributes are OBJECT types, resulted in a additional attributes being removed:

- Low Count attributes: Several attributes have a count that much lower than the amount of collisions. These will be not used.
- High Unique attributes: Several attributes have a large number of unique values, thus encompass many variables. For this exercise they also will be excluded. A maximum count of 15 variables per attribute is taken. Time/Location related attributes are not included with this.

Following this a improved data frame is constructed, figure 1. From this we now have a data frame with 189339 collisions and 18 attributes (1 binary class and 17 other), with 132221 class 1 and 57118 class 2 Severity entries.



```
Int64Index: 189339 entries, 0 to 194672
Data columns (total 18 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   SEVERITYCODE   189339 non-null  int64
 1   X              189339 non-null  float64
 2   Y              189339 non-null  float64
 3   LOCATION       189339 non-null  object
 4   COLLISIONTYPE  184582 non-null  object
 5   PERSONCOUNT    189339 non-null  int64
 6   PEDCOUNT       189339 non-null  int64
 7   PEDCYLCOUNT    189339 non-null  int64
 8   VEHCOUNT       189339 non-null  int64
 9   INCDATE        189339 non-null  object
 10  INCDTTM        189339 non-null  object
 11  JUNCTIONTYPE   185146 non-null  object
 12  SDOT_COLDESC   189339 non-null  object
 13  UNDERINFL      184602 non-null  object
 14  WEATHER        184414 non-null  object
 15  ROADCOND       184481 non-null  object
 16  LIGHTCOND      184327 non-null  object
 17  HITPARKEDCAR   189339 non-null  object
dtypes: float64(2), int64(5), object(11)
```

Fig 1. The cleaned data frame.

## 3. Methodology

A hypothesis is made that there will be a change in the amount of accidents in general and also that with time and improvement in car safety the amount of injuries (Severity Class 2) will go down faster than property damage (Severity Class 1).

As a first investigation, both classes are plotted against time being grouped per month, see figure 2 below.

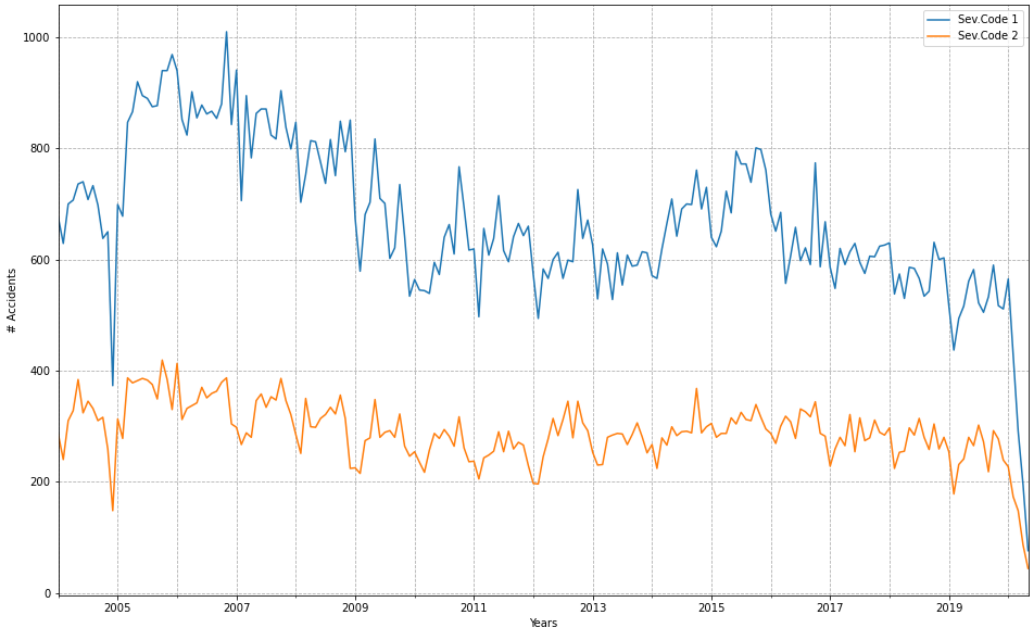


Figure 2, All reported cases for class 1 & 2

From this a first indication seems that the amount of injuries (Class 2) stays more stable compared to Class 1. See analysis for more details. Also a yearly pattern seems to emerge were the summer months have more accidents compared to the winter months. This seems the case for both classes. To further investigate the relation between both classes the ratio is determined over time, see figure 3.

Also here it seems that a yearly pattern emerges, perhaps not as clearly as in figure 2.

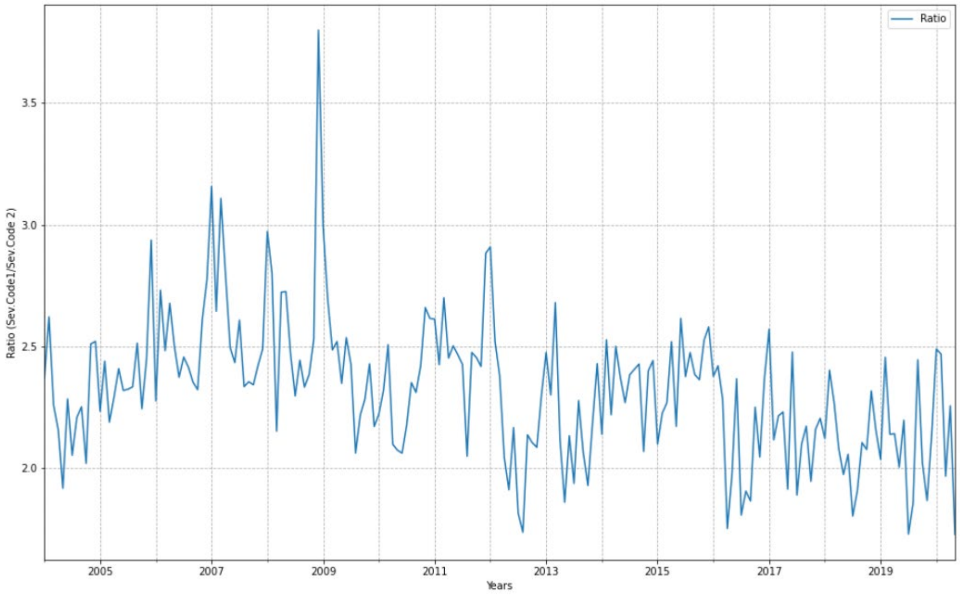


Figure 3: Ratio of (properties damage cases/ injury cases) over time.

## 4. Analysis

To see if indeed there's a decline over time for a linear regression is performed for # Sev.Code 1, # Sev.Code 2 over time. The details are show below in figure 4:

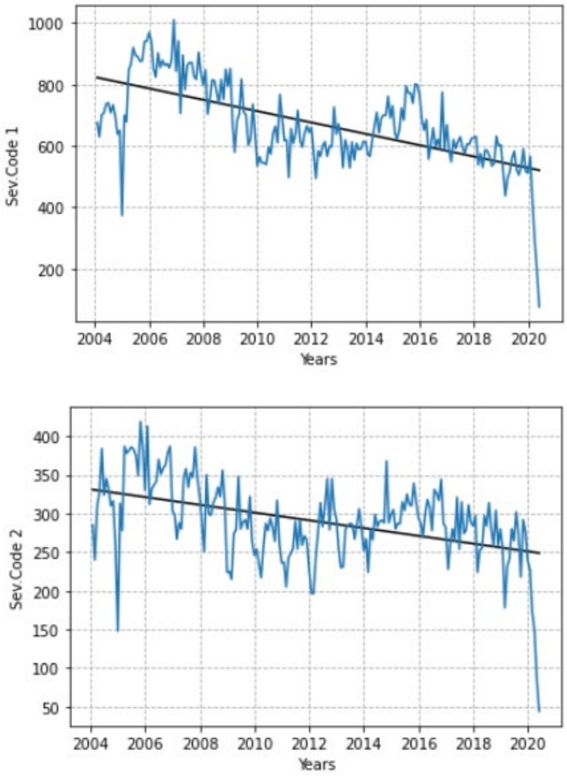


Fig. 4, linear regression of Sev.Code 1 (top) and Sev.Code 2 (bottom) over time.

From this this there is a clear trend showing that the number of cases is going down over time for both collateral damage (Sev.Code 1) and injury cases (Sev.Code 2)over the years. But it seems that the amount of collateral damage cases is going down faster than the injury cases.

To investigate this a linear regression is performed for the ratio of Sev.Code 1/Sev.Code 2 per month, which is shown in figure 5.

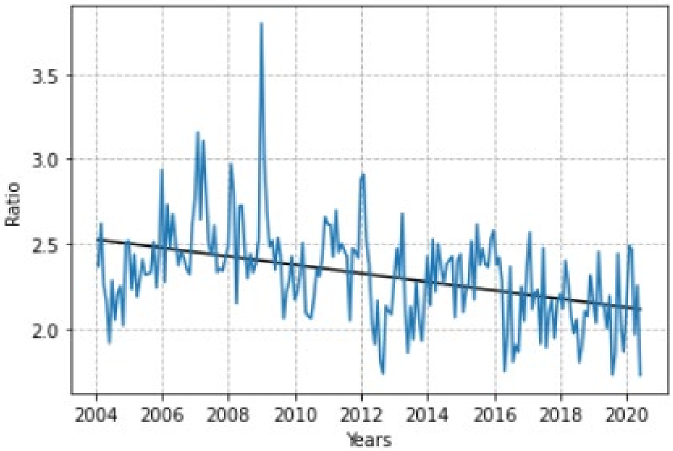


Fig. 5, linear regression of the ratio of Sev.Code 1/Sev.Code 2 over time.

Even though types of collision are decreasing, the small decline is visible in figure 5 indicates that the amount of injuries are increasing compared to the amount of collateral damage cases.

## 5. Results and Discussion

The choice was made to focus on the collision over time to see if a changes were notable in the total amount of collisions of both severity types. With the creation of the ratio attribute as described in section 4 provides in interesting addition to see changes of collisions over time.
A decrease in the total amount of accidents of both types of severity is detected but underlying patters are visible which need further investigation
For example, by investigating the yearly pattern more relevant data can be extracted and more precise modelling can be done. This was unfortunately not possible due to time constrains of the project.

A focus for further investigating should also be the OBJECT parameters. The can be changed by grouping into more suitable attributes, which can be explored to find certain parameters where specific severity types occur more frequently. For example see figure 6.

In figure 6, the heat map gives an indication of combination of parameters from both attributes are linked to a certain type of severity code or to a ratio of them. Eg. ‘Partly Clouded’ & ‘Wet’ has a blue value of 2, indication that with these conditions only severity code 2 accidents occur. Whilst ‘Partly Clouded’ and all other Road Condition parameters show a white value of 1.5 indicating a 1/1 ratio between code 1 and code 2 accidents.

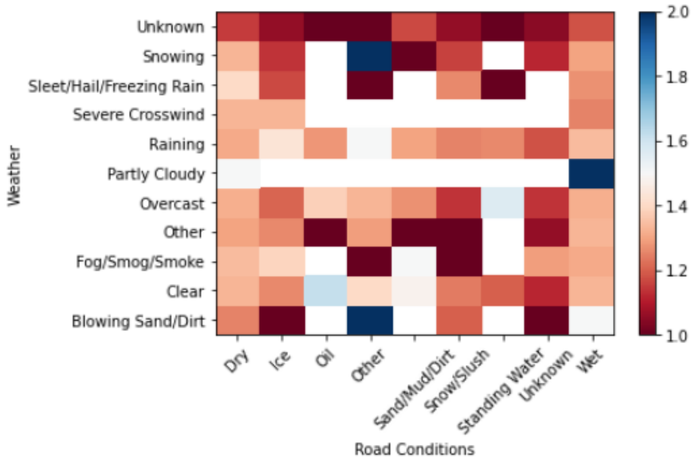


Fig. 6, A heatmap of Road Conditions and Weather Conditions. The color indicates the Sev.Code with Blue (2) being only injuries and Red being only collateral damage collisions (1). If a color is in between that means there are a ratio of both types.

A more details exploration is recommended.

## 6. Conclusion

Much more investigation of the created data frame as described in section 2 should be performed. Unfortunately to time constraints this wasn't possible.

It was shown that the amount of accidents decreased over time but that the amount of injury cases did not decrease as much as the amount of collateral damage cases. This was shown by creating a new parameter ['Ratio']. More detailed investigation of time aspects is recommended but for responders to accidents this might be an indicator that the focus should shift more injuries response than collateral damage response.

Further exploration by grouping the OBJECT parameters is recommended and an example is shown. Also the creation of additional parameters might hold additional information.

A final recommendation is to use the location parameters to investigate and visualize (with Folium) the area near Seattle that are more accident prone.