Predicting the Severity of Automobile Collisions

Matthew Levins

September 29, 2020

1 Introduction

1.1 Background

Public safety is of the utmost importance and finding ways to make our roads and neighbourhoods safer is a goal that continues to be sought after. One source of issue relates to traffic incidents causing harm to drivers and pedestrians. These traffic incidents vary drastically in their severity from small automobile damage to collisions that result in fatalities. Stakeholders such as governments, enforcement groups and the general public all have a common goal of reducing the chances of serious incidents occurring thus increasing the safety of drivers and pedestrians.

1.2 Problem

Therefore, the overall goal is to increase the safety of motorways and road networks by identifying the potential for a serious collision before the incident occurs. Understanding the trends and patterns that lead to these collisions can help government officials design and control the road system in a manner that is safer to the people who use it. It also allows current drivers to know if they need to be more cautious if their current conditions are identified as highly likely for a serious incident.

This problem requires the use of data science as there are many factors that contribute to collisions such as weather, traffic volume, the driver's state of mind and many more, plus the chances of unintuitive factors that may not be identified without the use of data science techniques. The goal is to identify the factors/patters that cause serious collisions, understand the magnitude of their individual impacts and develop a model that will predict the severity of a potential collision based on the current conditions.

2 Data

2.1 Data Source

To solve the problem of determining the severity of potential collisions, an effective dataset is required that has previous recorded incidents and their labeled severity along with the factors involved with that specific incident. Specifically, for this report, the dataset "Collisions – All Years" from SDOT Traffic Management Division will be used. This is a record of reported collisions in Seattle from 2004 to present day (September 2020).

2.2 Data Overview

The important attribute of this dataset is the labelled severity of each incident which has the following possibilities:

0 – Unknown

1 – Property Damage

2 – Injury

2b – Serious Injury

3 – Fatality

Initially, it is known that this rating system will be challenging to use due to the mix of alphanumeric characters, therefore this rating system will need to be modified to a scale of 1 to 4 and replacing the '2b' label with 3, the '3' label with 4, and eliminating the '0' label as the unknown severity incidents do not contribute to the main solution.

There are many attributes in this dataset to help predict the severity label, and with minor adjustments in formatting, they will become effective predictors. The list of attributes used for this problem are noted below:

- ADDRTYPE (Collision address type)
- INCDTTM (The date and time of the incident)
- JUNCTIONTYPE (Category of junction)
- UNDERINFL (If the driver was under the influence or not)
- WEATHER (Description of the weather at the time of collision)
- ROADCOND (Description of the road conditions)
- LIGHTCOND (Description of the lighting conditions)
- PEDROWNOTGRNT (Whether or not the pedestrian right of was granted)
- SPEEDING (If speeding was a factor)
- ST_COLCODE (Code relating to the description of the collision)
- HITPARKEDCAR (Whether or not the collision involved hitting a parked car)

For the attributes that are text descriptions (e.g. WEATHER), key words are isolated and used as the attribute for the prediction (e.g. "Rain"). Other formatting and data manipulation also help other attributes work effectively with the model development.

2.3 Data Cleaning

Before any analysis can be conducted on the data, it must be cleaned and prepared for such analysis to work effectively. Data cleaning consists of formatting the entries to be read and manipulated much easier in the later steps of this report. Upon initial inspection of the data, there were a few attributes that had many blank entries. It was found that attributes requiring a 'yes' or 'no' entry would only be filled in if there was a 'yes' response, therefore it can be concluded that the other blank entries were intended to be 'no' responses. The attributes that were affected was 'SPEEDING' and "PEDROWNOTGRNT' which were converted to numeric entries were '1' was a 'yes' and '0' was a 'no'.

After this process, there were still a few blank entries in the dataset, but not a lot, therefore any entry that had a blank response was removed from the overall dataset.

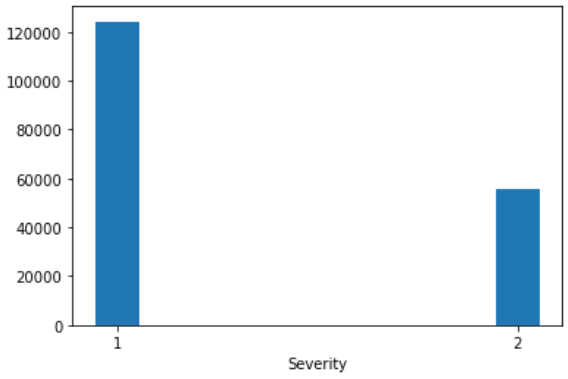
The next formatting task was associated with the 'INCDTTM' attribute that represents the time the accident occurred. This column was used to create new columns 'TIMEOFDAY' and 'DAYOFWEEK'. The time of the day column used the time from the original column and associated time intervals with 'Early Morning' (12am to 5am), 'Morning' (5am to 12pm), 'Afternoon' (12pm to 5pm), 'Evening' (5pm to 11pm). The original column was then dropped after the creation of the new columns that denoted the general time of the accident.

This cleaned dataset resulted in 180,068 observations with zero blank responses in the attributes chosen for the classification task.

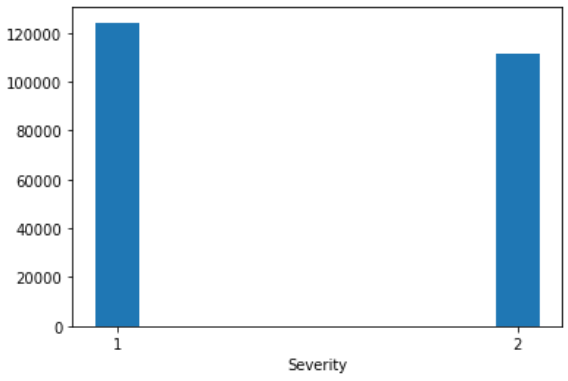
3 Exploratory Data Analysis (EDA)

3.1 Target Variable Inspection

The goal for this project was to predict the severity of a potential collision based on given factors. Therefore the first area of exploration was the target variable. As stated in the metadata, there are four levels of severity, however upon inspection, there were only two levels recorded in the entire dataset; level 1 and level 2. This converts the classification problem into a binary classification problem. The distribution of results is represented in the histogram below.



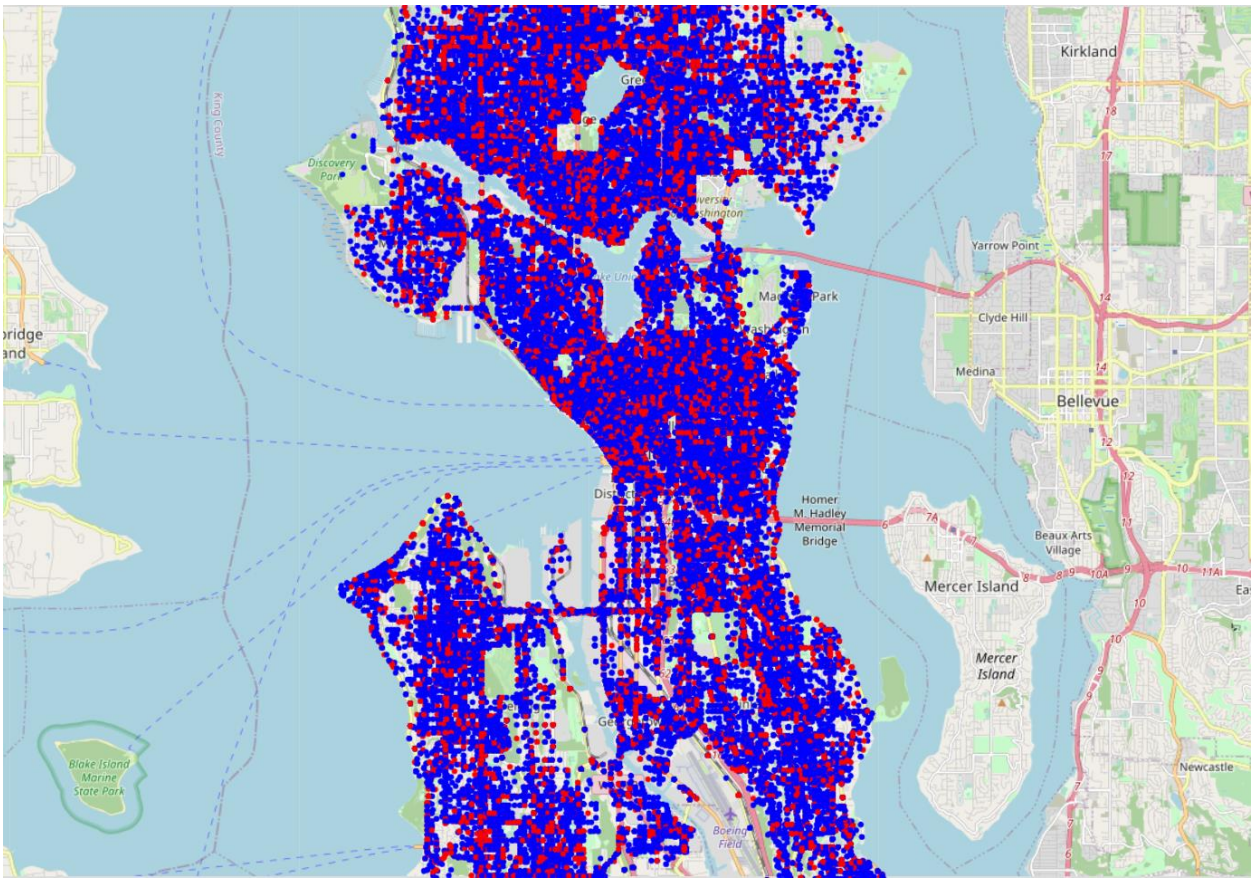
From this graph, it shows there are twice as many level 1 severities as level 2 which means less severe collisions occur much more often than severe ones. For equal comparison, the level 2 severity observations are doubled to have the same count as level 1. The results are shown below:



This result offers a more balanced dataset for model training and other EDA and allows attributes to be compared 1:1.

3.2 Visual Inspection

The first area of inspection for the data is the geographical location of the incidents and their related severities. If there is a cluster of high severity incidents in specific locations in Seattle, then this would be used for predictive purposes. The points were plotted over a map of Seattle for this visual inspection.



The image above plots the incidents from the dataset with red dots being a severity of '2' and blue dots having a severity of '1'. Upon visual inspection, there are no defined areas where higher severity collisions occur versus other locations. This eliminates the ability to predict a collision's severity from the geographical location.

3.3 Individual Predictor Relationships

3.3.1 Severity vs Junction Type/Address Type

The address type and junction type attributes both offer the same conclusion when comparing collision severity. It was discovered that "intersection related" incidents are typically more severe than other types of addresses or junctions. For address type, it was shown that a severe accident occurs 60% of the time when it is at a Junction. Similarly, the junction type reveals that a severe collision occurs 61% of the time when it is a junction related incident.

3.3.2 Severity vs Alcohol Impairment

When a collision occurs where a driver was deemed under the influence, a severe incident occurs slightly more frequently than not; the result was 55% of the time.

3.3.3 Severity vs Weather

The below table outlines the frequencies of each severity based on the weather condition. The severity of weather is negatively correlated to the severity of a collision. This is potentially due to the added caution and lower driving speeds when people are driving in bad weather conditions. Therefore if there is a collision, it wont be as severe.

SEVERITYCODE	1	2
WEATHER		
Partly Cloudy	0.250000	0.750000
Raining	0.491676	0.508324
Fog/Smog/Smoke	0.497954	0.502046
Clear	0.507693	0.492307
Overcast	0.515575	0.484425
Severe Crosswind	0.548387	0.451613
Blowing Sand/Dirt	0.580645	0.419355
Sleet/Hail/Freezing Rain	0.611511	0.388489
Snowing	0.681076	0.318924
Other	0.737470	0.262530
Unknown	0.876682	0.123318

3.3.4 Severity vs Road Condition

The below table outlines the frequencies of each severity based on the road condition. Road conditions related to harsh weather also reduce the severity of collisions. Severe incidents occur with less than ideal road conditions at unexpected times (e.g. oil on the road). This relates back to the conclusion made about weather as drivers are more cautious when conditions are unfavourable, but this caution is not given all the time which lead to severe conditions due to unexpected road conditions.

SEVERITYCODE	1	2
ROADCOND		
Oil	0.420290	0.579710
Sand/Mud/Dirt	0.476190	0.523810
Other	0.493506	0.506494
Wet	0.497414	0.502586
Dry	0.508743	0.491257
Standing Water	0.584000	0.416000
Ice	0.629474	0.370526
Snow/Slush	0.709507	0.290493
Unknown	0.885549	0.114451

3.3.5 Severity vs Light Conditions

The amount of light in the environment is negatively correlated with the severity of the collision. The darker the conditions, the more cautious the driver thus resulting in lowering the chance for a severe collision. See the table below for the results from this relationship.

SEVERITYCODE	1	2
LIGHTCOND		
Dark - Unknown Lighting	0.466667	0.533333
Dawn	0.494994	0.505006
Daylight	0.496556	0.503444
Dusk	0.499540	0.500460
Dark - Street Lights On	0.535726	0.464274
Dark - Street Lights Off	0.576011	0.423989
Dark - No Street Lights	0.633641	0.366359
Other	0.645933	0.354067
Unknown	0.895335	0.104665

3.3.6 Severity vs Pedestrian's Right of Way

When a pedestrian is involved with the collision, there is a much higher chance for a severe incident. This is evident with when investigating the collisions where the pedestrian's right of way was not granted. It is 18 times more likely for this type of incident to be a level 2 severity than a level 1.

3.3.7 Severity vs Speeding

When investigating the relationship between severity and speeding, the results showed that level 2 severity incidents occur 55% of the time.

3.3.8 Severity vs Day of The Week

The following table shows the frequency of each level of severity for the different days of the week. A conclusion from this result relates to collisions on the weekend versus a weekday. There is a tendency for less severe collisions on the weekend compared to weekdays.

SEVERITYCODE	1	2
DAYOFTHEWEEK		
Thursday	0.514614	0.485386
Tuesday	0.518030	0.481970
Wednesday	0.519133	0.480867
Monday	0.519947	0.480053
Friday	0.531125	0.468875
Saturday	0.536265	0.463735
Sunday	0.555040	0.444960

3.3.9 Severity vs Time of Day

The time of day does relate to the severity of collisions and this relationship seems to correlate with the lighting conditions relationship. In the early morning and evening where the environment would be darker, the collisions are less severe than the morning and afternoon collisions. The table below shows the split between the severity levels based on the time of day.

SEVERITYCODE	1	2
TIMEOFDAY		
Afternoon	0.498278	0.501722
Morning	0.525797	0.474203
Evening	0.527289	0.472711
Early Morning	0.569526	0.430474

4 Predictive Modeling

The nature of this problem is a binary classification problem that can take advantage of multiple models. The models must use the attributes discussed earlier in this report to predict whether the severity of the collision will be level 1 or level 2. The final dataset that was used was a one-hot-encoded data frame that outlines which conditions are present and has the corresponding label of severity.

Id	Attribute	Count
0	SEVERITYCODE	235877
1	INTERSECTIONRELATED	235877
2	BLOCKNOTINTERSECTIONRELATED	235877
3	UNDERINFL	235877
4	Raining	235877
5	Fog/Smog/Smoke	235877
6	Blowing Sand/Dirt	235877
7	Sleet/Hail/Freezing Rain	235877
8	Snowing	235877
9	Oil	235877
10	Sand/Mud/Dirt	235877
11	Wet	235877
12	Standing Water	235877
13	Ice	235877
14	Snow/Slush	235877
15	DarkNoLights	235877
16	PEDROWNOTGRNT	235877
17	SPEEDING	235877
18	WEEKEND	235877
19	Afternoon	235877
20	Early Morning	235877

4.1 Classification Models

The chosen models for this problem were Decision Trees, KNN and Logistic Regression. These models were trained on the same data that was split from the original data leaving 30% for model testing. Each model had at least one hyperparameter that required exploration; therefore, the models were run multiple times to determine the best hyperparameters. The results from the test data for each model s shown below where the accuracy scores are based on the performance of the 30% test data.

Model	Accuracy
Decision Tree	61.6%
KNN	61.2%
Logistic Regression	61.4%
Logistic Regression (Probability Prediction)	64.8%

All accuracies were similar between models however their accuracy levels require improvement to be a more effective classifier. It was also noted that KNN required much high computation time to perform similar to the other models resulting in the decision to remove it from further investigation.

4.2 Principal Component Analysis

To potentially increase the performance of the classifier models, PCA was conducted on the data from before. The arbitrary dimension reduction for PCA was to reduce the 20 attributes to 5 new inputs for

the model construction. The decision to reduce dimensionality was from the fact that some of the attributes in the original data were seemingly dependent such as lighting conditions and time of day as well as road conditions and weather. By reducing the dimension of the data, the hope was to have more impactful inputs for the model to train on. After PCA was conducted, the models were trained again producing the following results.

Model	Accuracy
PCA Decision Tree	61.5%
PCA Logistic Regression	60.8%
PCA Logistic Regression (Probability Prediction)	66.4%

The only improvement from PCA was the performance of the probability prediction in the Logistic Regression Model.

5 Summary of Results and Discussion

Results from the EDA are listed below and provide a qualitative perspective on the data:

1. Collisions related to intersections are more likely of being a severe collision
2. A collision due to impairment is more likely of being a severe collision
3. The severity of weather is negatively correlated to the severity of a collision. This is potentially due to the added caution and lower driving speeds when people are driving in bad weather conditions. Therefore if there is a collision, it wont be as severe.
4. Road conditions related to harsh weather also reduce the severity of collisions. Severe road conditions occur with less than idea road conditions at unexpected times (e.g. oil on the road)
5. The amount of light in the environment is also negatively correlated with the severity of the collision. The darker the conditions, the more cautious the driver and potentially a less populated area resulting in less severe collisions.
6. Collisions involving a lack of granting a pedestrian's right of way is highly likely to be a severe collision.
7. Speeding causes more severe collisions however, it is not as drastically proportioned as one might expect.
8. Weekend collisions are slightly less severe than weekday collisions
9. Collisions in the afternoon are more severe than collisions in the early morning. This is probably related to the lighting conditions as well.

For this report, five different models were trained and validated. These models attempted to conduct binary classification based on the given inputs that were constructed in the Data Preparation phase of the analysis. The general accuracies from these models are below a desired level however they offer the preliminary steps to construct a better model in the future.

All models provided an average of roughly 61% accuracy which shows that the data itself needs improving and more in-depth EDA to potentially uncover more attributes for the prediction models. Furthermore, it was shown that PCA did not improve the models' accuracy thus proving the need for better EDA.

6 Conclusion

The results for constructing a predictive model to gauge whether a potential collision will be severe or not was not as effective as hoped in the beginning of the project. The EDA was helpful in finding correlations between different attributes and the severity of the collisions however these correlations were not as effective when using the predictive model. The linear regression model would be the most effective to use as a probability model rather than a definitive classification model. This is shown with the log-loss accuracy of ~65%, which is higher than the other models. The output of the model would give the probability of a severe collision occurring rather than a definitive yes or no, which the driver can take into consideration if the model was implemented live.

It was interesting to see that severe driving conditions lead to collisions of less severity which means the caution that drivers take does contribute to a safer roadway. Areas that need focus are collisions when pedestrians are involved and during unexpected road conditions like oil and mud on the road.