Reducing Injury Rates from Seattle Collisions

Introduction

The Seattle Department of Transport (SDOT) is a municipal agency in Seattle, Washington responsible for the maintenance of the city's transportation systems, including roads, bridges, and public transportation. An obvious goal of SDOT is the protection from injury of both the property and person of users of the Seattle transportation system. This general goal suggests a range of more specific business problems for the organisation to consider. One of the most important among these can be stated as follows:

"What policies, regulations, and public communication strategies can best serve to reduce the risk of personal injury to users of the public road system?"

This report seeks to address this problem directly, through the utilisation of a public database of traffic accidents produced by SDOT. The report is intended to be widely disseminated, both publicly and within SDOT, however it is primarily directed to the key policy making staff who are in a position to effect change within the current suite of policies, rules, regulations, and public communication strategies.

Specifically, the intention is to deepen understanding of the specific factors that increase risks of personal injury resulting from traffic accidents through the development of a predictive model. It is hoped that insights gained from the model may enable the existing suite of policy measures and traffic rules and regulations to be improved with the goal of decreasing the incidence of personal injury from traffic accidents in Seattle.

Data

The dataset used in this analysis was acquired through the Seattle Open Data Portal and can be accessed through the following link:

https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions/

At the time the data was extracted for analysis it contained a total of 221,389 records representing collisions occurring within Seattle City between the dates of 06-10-2003 and 05-09-2020.

The raw dataset contained a total of 40 data fields for each collision. Consideration was given to which subset of these 40 fields would provide the most information-rich feature set for the predictive model. Considerations included the degree of overlap between fields, the data quality and degree of missing data, and the applicability of each field to the specific business problem.

Collisions with either missing data or where critical information was registered as "unknown" were omitted.

The final feature set includes the following:

1) SEVERITYCODE (Target)

The raw data contains 5 possible severity codes as shown below. In order to better address the business problem these codes were mapped to a simple binary classification representing those collisions resulting in personal injury (1) and those that did not (0). The mappings are also shown below. Collisions with an unknown severity code were discarded

- 3 – fatality => 1
- 2b – serious injury => 1
- 2 – injury => 1
- 1 – property damage => 0
- 0 – unknown => discarded

The resulting dataset contained 62,198 collisions with injury and 137,596 collisions with no injury. This field represents the target variable or label used for the predictive model.

2) INATTENTIONIND

This field contains "Y" when the collision was due to inattention and "NaN" otherwise. These classes were mapped to 1 and 0 respectively for modelling. All else being equal it could be expected that driver inattention may increase the likelihood of both collisions in general and collisions resulting in injury.

The resulting dataset contained 28, 710 instances of inattention and 144, 495 instances where inattention was not a factor

3) UNDERINFL

This field contained mixed data. Either a "Y" or a "1" represented instances where a driver was under the influence of drugs or alcohol and "N" or "0" for instances when this was not the case. These mixed classes were mapped to 1 and 0 respectively for modelling. All else being equal it could be expected that the presence of drugs or alcohol may increase the likelihood of both collisions in general and collisions resulting in injury.

The resulting dataset contained 9,629 instances where drugs or alcohol were a factor and 185,550 instances where they were not a factor

4) SPEEDING

This is another binary field with a "Y" indicating speeding was a factor in the collision and "NaN" otherwise. As with INATTENTION these classes were mapped to 1 and 0 respectively. All else being equal it could be expected speeding may increase the likelihood of both collisions in general and collisions resulting in injury.

The dataset contained 9,628 instances where speeding was a factor and 163,577 where it was not.

5) PEDCOUNT and PEDCYLCOUNT

These two fields represent the number of pedestrians or cyclists involved in the collision respectively. These fields were aggregated into a single binary feature representing whether or not either cyclists or pedestrians were involved. This created a more robust single indicator for the presence of non-vehicle participants in a collision.

Due to the fact that pedestrians and cyclists are generally far less protected than participants in vehicles it seems likely that injuries may be more likely were they are involved in a collision.

The engineered binary feature contained 13,965 instances where non-vehicle persons were involved and 207,424 where they were not.

6) JUNCTIONTYPE

This field indicates 6 different types of location for the collision. To facilitate modelling these classes were mapped to binary indicatory variables using one-hot encoding.

All else being equal it could be assumed that intersections carry more risk for accidents and injury than mid-block locations. This is due to both the requirement for drivers to make more complex decisions at intersections and also because any collision at an intersection is more likely to involve vehicles travelling in different directions and hence potentially higher overall impact.

The table below shows the relative frequency of the 6 classes

	JUNCTIONTYPE
Mid-Block (not related to intersection)	77646
At Intersection (intersection related)	61872
Mid-Block (but intersection related)	21522
Driveway Junction	10177
At Intersection (but not related to intersection)	1826
Ramp Junction	162

7) WEATHER

This feature indicates the prevailing weather conditions at the time of the collision. To facilitate modelling these classes were mapped to binary indicator

variables using one-hot encoding.

The influence of weather is likely to be quite subtle and potentially dependent on other factors in combination. While it could be argued that difficult conditions (eg rain, snow) might increase risks of accident, these risks are often mitigated by the tendency for vehicles to travel more slowly during difficult conditions. This may lead to a more muted impact on injury prevalence.

The table below shows the relative frequency of the 10 classes.

	WEATHER
Clear	110797
Raining	33192
Overcast	27381
Snowing	830
Fog/Smog/Smoke	552
Other	261
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	43
Severe Crosswind	26
Partly Cloudy	10

8) ROADCOND

This feature indicates the road conditions at the time of the collision. To facilitate modelling these classes were mapped to binary indicatory variables using one-hot encoding.

As with weather there are competing influences on risk of injury due to more difficult conditions being offset by lower speeds in many instances.

The table below shows the relative frequency of the 8 classes.

	ROADCOND
Dry	123830
Wet	47088
Ice	1104
Snow/Slush	845
Other	109
Standing Water	108
Sand/Mud/Dirt	63
Oil	58

9) LIGHTCOND

This feature indicates the light conditions at the time of the collision. To facilitate modelling these classes were mapped to binary indicatory variables using one-hot encoding.

As with weather and road conditions there are competing influences on risk of injury due to more difficult conditions being offset by lower speeds in many instances.

The table below shows the relative frequency of the 8 classes.

	LIGHTCOND
Daylight	114499
Dark - Street Lights On	47725
Dusk	5750
Dawn	2487
Dark - No Street Lights	1400
Dark - Street Lights Off	1130
Other	193
Dark - Unknown Lighting	21

Example collision

The following table describes an example collision using the raw data contained in the above fields prior to engineering into binary features. Not

SEVERITYCODE	2
JUNCTIONTYPE	At Intersection (intersection related)
INATTENTIONIND	NaN
UNDERINFL	N
WEATHER	Overcast
ROADCOND	Wet
LIGHTCOND	Dark - Street Lights On
SPEEDING	NaN
PEDCOUNT	0
PEDCYLCOUNT	0

Data Summary

Each of the features described could be expected to have some influence on the likelihood of an accident occurring and on the likelihood of an accident leading to injury, as opposed to purely property damage. Analysis of the individual features described above and the development of a predictive model for collision injury should assist in providing SDOT policy makers with a more informed basis for improving existing policy.