Reducing Injury Rates from Seattle Collisions

Toby Masters

September 20, 2020

1. 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.

2. Data acquisition and cleaning

2.1 Data source

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.

2.2 Data cleaning and feature extraction

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.

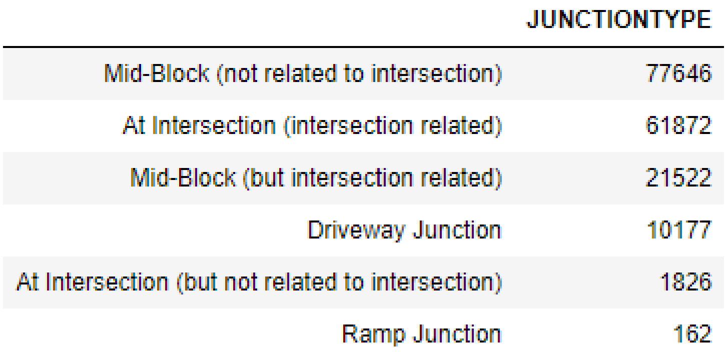
The engineered binary feature contained 13,965 instances where non-vehicle persons were involved and 207,424 where they were not.

6) JUNCTIONTYPE / ADDRTYPE

The JUNCTIONTYPE 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

In the final analysis the JUNCTIONTYPE feature was replaced by ADDRTYPE which indicates simply whether the collision occurred at an intersection, within a block or in an alley.

The small number of alley based collisions were discarded and the ADDRTYPE was encoded into a binary variable with 1 representing intersection.

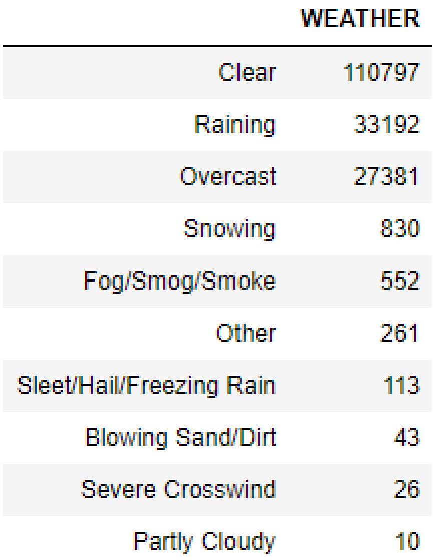
7) WEATHER

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.



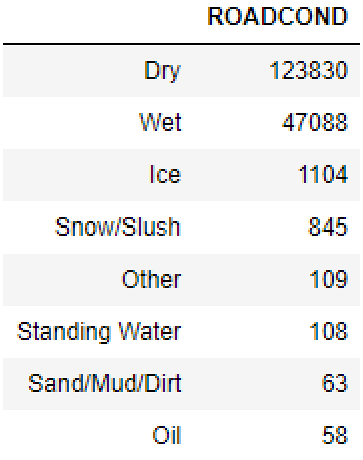
	WEATHER
Clear	110797
Raining	33192
Overcast	27381
Snowing	830
Fog/Smog/Smoke	552
Other	261
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	43
Severe Crosswind	26
Partly Cloudy	10

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.



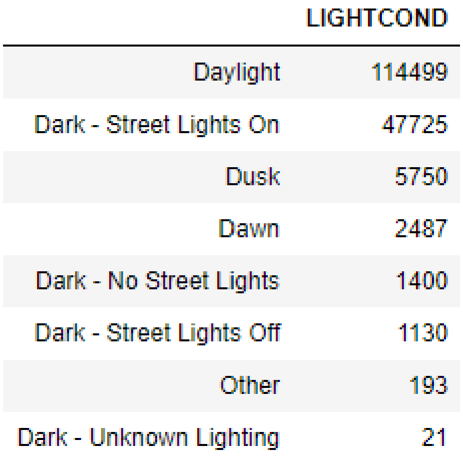
	ROADCOND
Dry	123830
Wet	47088
Ice	1104
Snow/Slush	845
Other	109
Standing Water	108
Sand/Mud/Dirt	63
Oil	58

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.

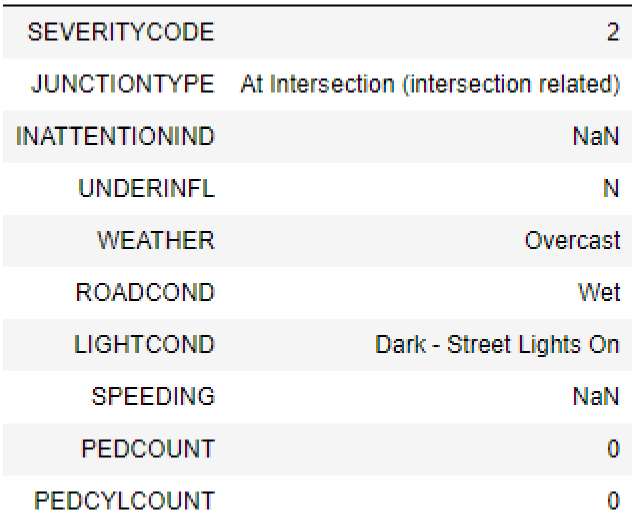
The table below shows the relative frequency of the 8 classes.



	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

2.3 Example collision

The following table describes an example collision using the raw data contained in the above fields prior to engineering into binary features.



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

2.4 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.

3. Methodology

3.1 Target Variable

The target variable for the analysis was the SEVERITYCODE as described in the Data section. The raw data was mapped to a single binary variable indicating whether or not a collision resulted in injury.

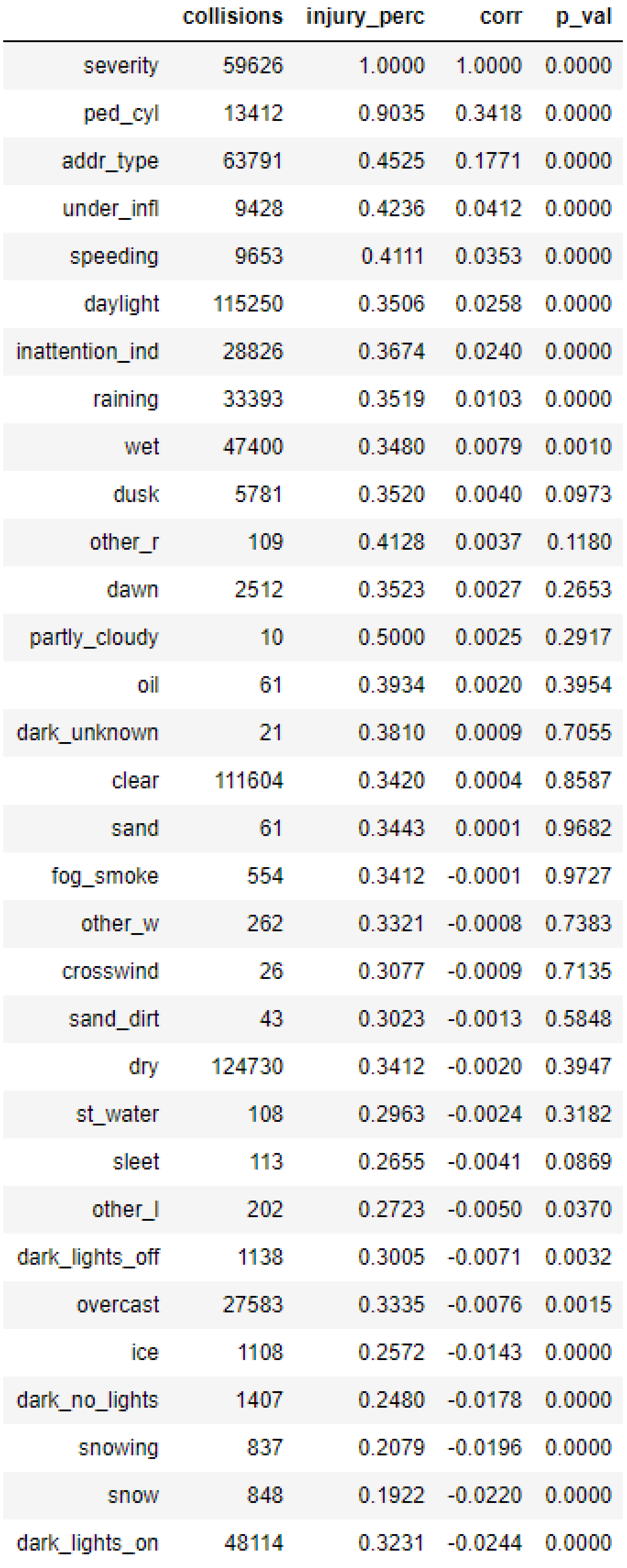
As described in the Data section the target variable has unbalanced labels, with 62,198 collisions involving injury and 137,596 involving no injury. After cleaning the full feature set to remove incomplete data this was reduced to 59,498 collisions with injury and 113,707 collisions without injury, or roughly a 1/3 - 2/3 ratio.

In order to balance the labels both upsampling and downsampling methodologies were tested to avoid introducing bias into the model. While both methodologies yielded similar overall results downsampling was preferred as there was sufficient data available to avoid the duplication of collisions introduced through an upsampling approach.

3.2 Exploratory data analysis

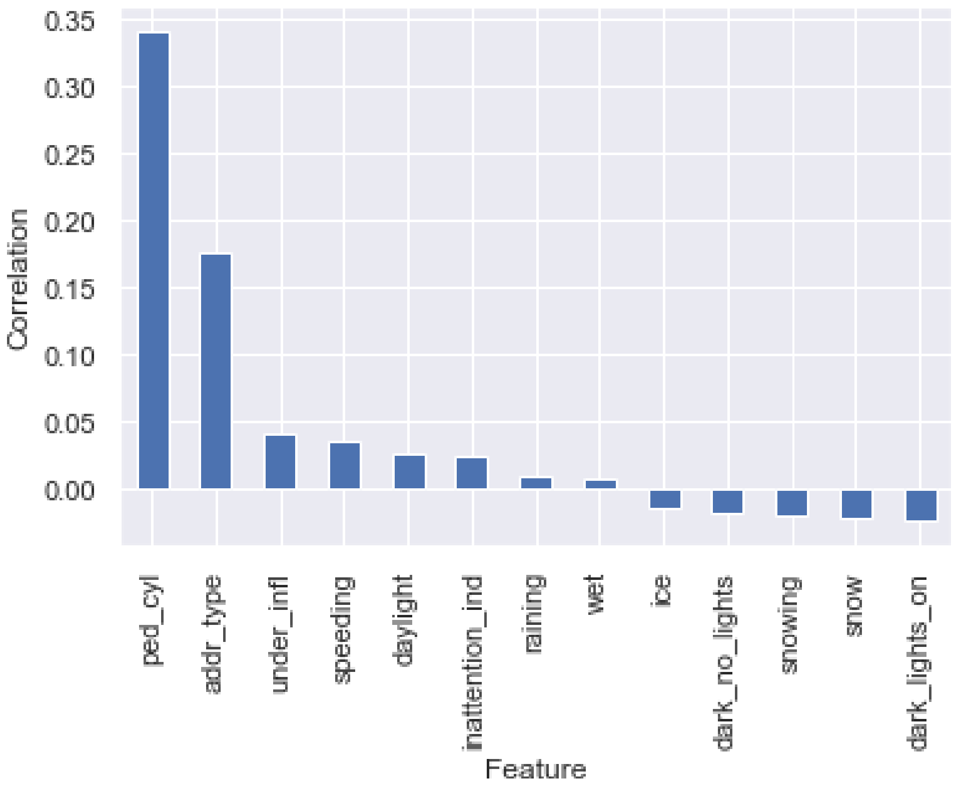
After re-engineering the features into binary variables and following the one-hot encoding procedure described in the Data section the entire feature set consists of binary categorical variables. In many cases the data contained in these features is sparse making visual and statistical inference more difficult. For example features that are associated with a high injury rate may be prima facie important factors but if the number of collision instances exhibiting this feature is low then we may be more sceptical that this result may be generalisable.

The table below shows summary statistics for each feature including the number of positive instances, the proportion of positive instances associated with injury, the Pearson correlation and the associated p-value.



	collisions	injury_perc	corr	p_val
severity	59626	1.0000	1.0000	0.0000
ped_cyl	13412	0.9035	0.3418	0.0000
addr_type	63791	0.4525	0.1771	0.0000
under_infl	9428	0.4236	0.0412	0.0000
speeding	9653	0.4111	0.0353	0.0000
daylight	115250	0.3506	0.0258	0.0000
inattention_ind	28826	0.3674	0.0240	0.0000
raining	33393	0.3519	0.0103	0.0000
wet	47400	0.3480	0.0079	0.0010
dusk	5781	0.3520	0.0040	0.0973
other_r	109	0.4128	0.0037	0.1180
dawn	2512	0.3523	0.0027	0.2653
partly_cloudy	10	0.5000	0.0025	0.2917
oil	61	0.3934	0.0020	0.3954
dark_unknown	21	0.3810	0.0009	0.7055
clear	111604	0.3420	0.0004	0.8587
sand	61	0.3443	0.0001	0.9682
fog_smoke	554	0.3412	-0.0001	0.9727
other_w	262	0.3321	-0.0008	0.7383
crosswind	26	0.3077	-0.0009	0.7135
sand_dirt	43	0.3023	-0.0013	0.5848
dry	124730	0.3412	-0.0020	0.3947
st_water	108	0.2963	-0.0024	0.3182
sleet	113	0.2655	-0.0041	0.0869
other_l	202	0.2723	-0.0050	0.0370
dark_lights_off	1138	0.3005	-0.0071	0.0032
overcast	27583	0.3335	-0.0076	0.0015
ice	1108	0.2572	-0.0143	0.0000
dark_no_lights	1407	0.2480	-0.0178	0.0000
snowing	837	0.2079	-0.0196	0.0000
snow	848	0.1922	-0.0220	0.0000
dark_lights_on	48114	0.3231	-0.0244	0.0000

The bar chart below shows the subset of features with statistically significant Pearson correlations (p-value < 0.001).



A few observations can be made:

- The pedestrian/cyclist feature shows significant positive correlation with injury. This is not surprising given the relative vulnerability of pedestrians / cyclists versus motor vehicle passengers
- Intersections (addr_type=1) are correlated with a higher incidence of injury. This could be a function of more complex driver environments and increased impacts from multidirectional vehicles.
- Drugs/alcohol, speeding and inattention show a positive correlation to injury consistent with expectations suggesting increased driver education and awareness may be valuable
- The positive correlation to daylight is interesting. One would assume that daylight conditions are safer than other lighting conditions. It is possible this leads to greater average speeds and hence injury risk.
- Interestingly all of the statistically significant features with negative correlation to injury (albeit very low negative correlations) involve poor conditions (ice, snow, darkness). The negative correlation may suggest that while these conditions are

associated with collisions, the lower speeds employed by drivers in these conditions may be leading to lower incidence of injury.

3.3 Machine learning models

A range of machine learning (ML) algorithms exist for producing predictive classification models. The primary focus of this analysis has been Decision Trees and Logistic Regression as these can perform well with sparse datasets both with respect to model results and calculation time. Support Vector Machines and K Nearest Neighbours models were also investigated for comparison although calculation time considerably slower with no increase in performance.

One particular benefit of the Decision Tree model is the relative ease of interpreting the results. This higher degree of interpretability will be helpful in drawing conclusions around potential policy shifts in future efforts to reduce injury rates.

With the need for interpretability and clarity in mind the complexity of the initial feature set was reduced slightly for a second iteration of model training. Specifically the JUNCTIONTYPE feature with six original classes was replaced by a single binary feature representing whether the collision occurred at an intersection or mid-block. This change led a more easily interpretable decision tree model with more obvious implications for policy.

The data was split into separate training and test sets and all models were optimised for test set performance according to their key sensitivity parameters.

4. Results

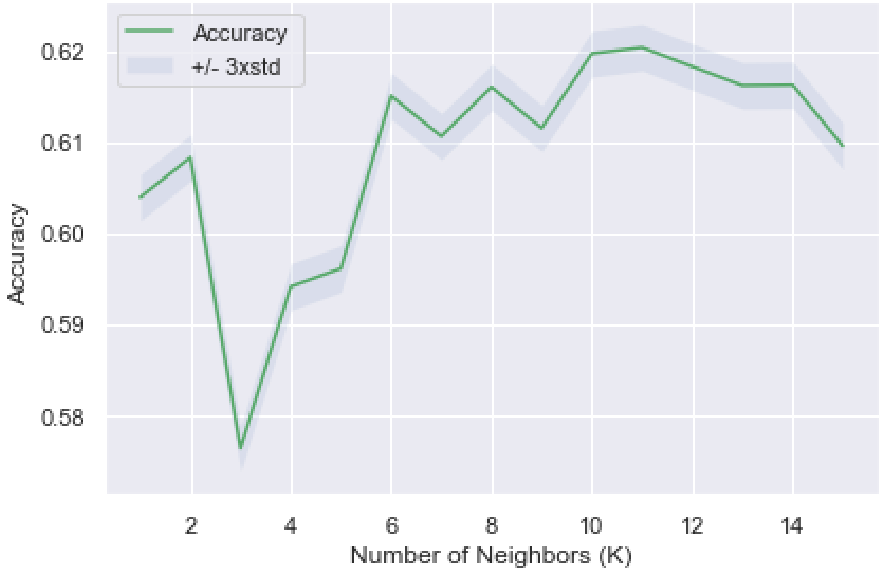
Comparative results for six ML models are presented in the table below. The results for the unbalanced target set using a decision tree model are shown for interest only. While this model had the highest accuracy and F1 score we can see that this is at the expense of very poor false negative performance. Essentially the biased model is overpredicting injury significantly. Given the raw incidence of injury is around 35% we know that a purely naïve model could expect to generate 65% accuracy on the unbalanced labels. On this basis 72.1% accuracy is not a dramatic improvement over a naïve model.

After balancing the labels we can obtain a more unbiased view of the predictive ability of the models. The models all perform with similar accuracy and F1 score although with somewhat different distributions within the confusion matrices. The calculation performance of the decision tree and logistic regression models was significantly better than SVM and KNN. The best performance for each metric is highlighted in red (excluding the unbalanced decision tree).

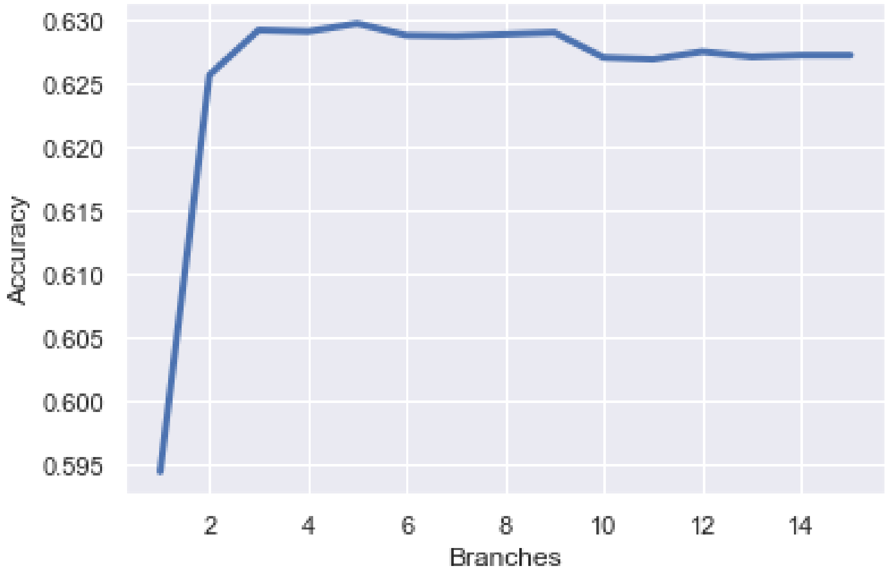
	Unb. Dec Tree (4)	Decision Tree (3)	Decision Tree (5)	SVM	KNN	Logistic Regression
Accuracy	0.721	0.629	0.630	0.629	0.620	0.628
True Positives	0.984	0.657	0.679	0.657	0.722	0.691
False Positives	0.016	0.343	0.321	0.343	0.278	0.309
True Negatives	0.215	0.602	0.581	0.602	0.480	0.565
False Negatives	0.785	0.398	0.419	0.398	0.520	0.435
F1 Score	0.659	0.629	0.629	0.629	0.616	0.626

The SVM was optimised for kernel and C parameter although the parameter sensitivity was not overly significant. The SVM performance metrics were actually identical to the three branch decision trees' results albeit with significantly lower calculation efficiency. Likewise the logistic regression model had low sensitivity to the C parameter and had similar performance metrics although was relatively strong in true positives at the expense of slightly lower true negatives.

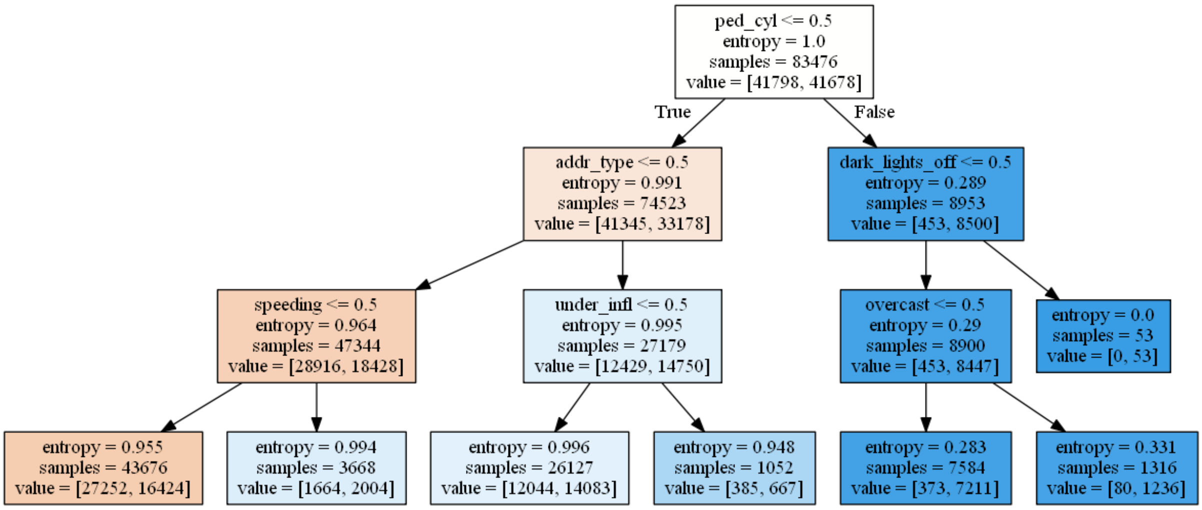
The KNN model was slow to calculate given the feature set size and sparse nature of the data but ultimately performed similarly to the other models (albeit with a less balanced confusion matrix) at higher levels of K, as can be seen in the plot below.



Most of the predictive ability in the decision tree model was acquired by the third branch as the sensitivity plot below indicates.



While the five branch decision tree showed slightly better overall accuracy, the three branch tree is easier to interpret in terms of policy implications and is produced below.



As already implied by the inferential statistical analysis the existence of pedestrian/cyclist involvement in a collision is highly predictive of injury. Not surprisingly in cases where pedestrians/cyclists were involved and lack of lighting or overcast conditions were also present there is further increased risk of injury. Presumably because pedestrians are more difficult for drivers to see in the dark.

For collisions not involving pedestrians/cyclists the model confirms the significance of intersections as being more predictive of injury than collisions occurring within blocks. Interestingly the existence of drivers under the influence of drugs or alcohol was the most significant predictor of injury for collisions occurring within intersections while speeding was more significant for collisions occurring within blocks.

The observations suggest potential policy directions that will be outlined in the conclusion section.

5. Conclusion

The SDOT collisions database has enabled us to distil a number of key findings from our exploratory analysis and predictive model development. These findings should assist in directing future SDOT policy development with the goal of reducing the incidence of injury from collisions.

Key conclusions and implied policy responses are summarised below:

- Pedestrians remain vulnerable on our roads, particularly in conditions where light conditions are poor. Further research should look to highlight areas of particularly pedestrian vulnerability and consider the possibility of improving artificial lighting in these areas to assist drivers in identifying pedestrians early. Variable speed limits could also be considered based on time of day where pedestrian activity is most significant. Further research into time of day effects would be useful in supporting this proposal.
- Collisions occurring at intersections are more likely to produce injuries than those occurring within blocks. Intersections require increased awareness from drivers and carry additional risk of high impact due to the potential for multi-direction collisions. Consideration should be given to finding opportunities to simplify driver decisions at intersections. A further review of location data to highlight any hotspots may assist in better targeting a response.
- Drugs and alcohol remain a significant cause of injury on our roads. Existing policies around public awareness campaigns and infringement penalties should be reviewed in light of the evidence of continued impact on injury rates.
- Mid-block collisions are more likely to lead to injury where speeding is a factor. Speed cameras, and speed indicators have proven to be effective in influencing driver behaviour. A review should be conducted to identify any further opportunities to use these tools to influence drivers and reduce the incidence of speeding where risks are measurably highest.
- Surprisingly, poor weather and lighting conditions are not a significant factor in injury rates on our roads, with the exception of collisions involving pedestrians as noted above. Presumably drivers are managing the increased risk under such conditions by reducing speeds and this is having a favourable affect injury rates despite collision risk.