Harnessing Data Science to inform SDOT injury prevention policy improvements

TOBY MASTERS

SEPTEMBER 2020

“As data scientists, our job is to extract signal from noise.”

- DANIEL TUNKELANG

Identifying the drivers of collision injury can assist in improving SDOT policy

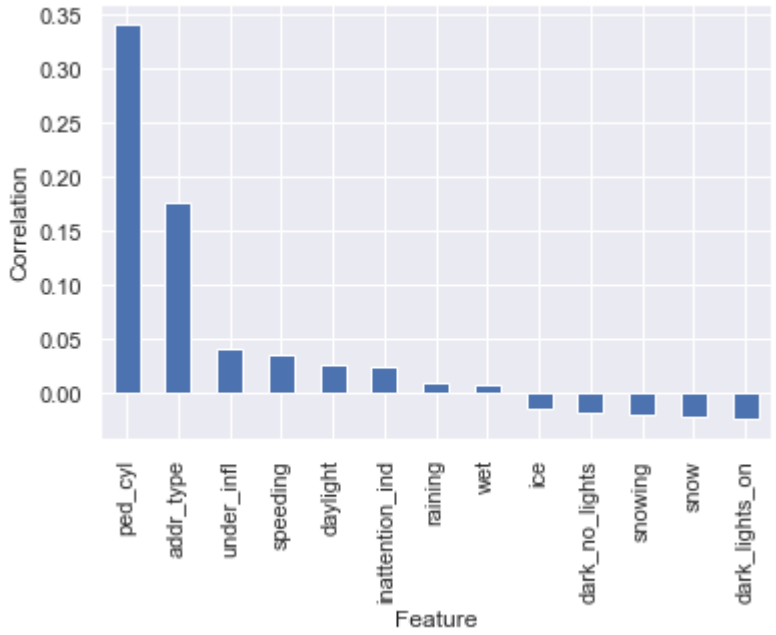
- Injury rates remain too high on Seattle roads despite improvements over recent years
- Applying data science to collisions data offers the potential to better understand the key drivers of injury and thus better inform SDOT policy improvements
- SDOT injury prevention policy is implemented through a combination of rules and regulations, traffic signs, signals, cameras, and through public communication and awareness campaigns
- Each of these policy delivery mechanisms has the potential to be made more effective by applying the insights garnered through data science – extracting signal from noise

Data Acquisition and Cleaning

- 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/
- 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
- The target variable was reduced to a binary variable indicating whether or not a collision resulted in injury
- Of the remaining data fields 8 features were extracted and cleaned. This was expanded to 31 binary features via one-hot encoding
- After downsampling the data to balance target labels a total of 119,252 collisions were used for modelling

Data exploration and inferential analysis

- Initial exploration of the data revealed a subset of features exhibiting statistically significant correlation to collision injuries
- Several observations can be made
  - Pedestrians/cyclist are particularly vulnerable
  - Intersections appear to offer greater risk of injury than mid-block collisions
  - Drugs/alcohol, speed and driver inattention are also correlated to injury
  - Surprisingly poor conditions such as snow and ice are negatively correlated to injury suggesting lower speeds may be a factor

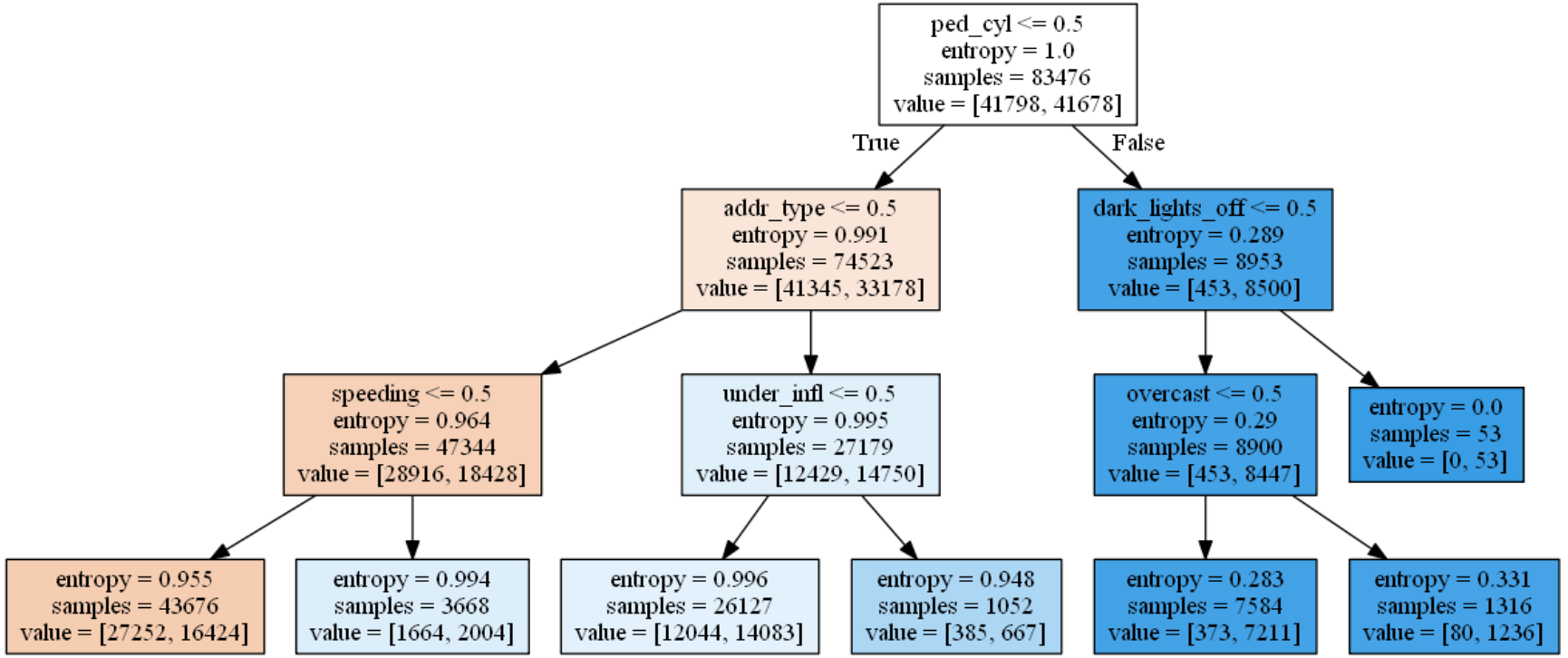


Model Results

- A range of machine learning models were analysed including Decision Trees, K nearest neighbour, Support Vector Machines and Logistic Regression
- SVM and KNN were computationally slow although SVM statistical metrics were identical to a 3 branch Decision Tree
- Performance was similar across models with moderate difference in confusion matrix balance
- Decision Trees potentially the most useful for informing policy changes

	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

3 branch decision tree offered the best interpretability



Conclusion

- 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 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.