



Predicting accident severity

Jonathan Koh



Introduction



➤ Background

- Road accidents happen everyday and there may be certain common factors that contribute to an increased likelihood of getting into an accident that is severe.
- By being able to predict the severity of an accident the police may be able to get a quick sense on whether the accident is likely to be severe and can request for resources like ambulance if an accident is likely to involve injuries

➤ Problem

- To predict the severity of any car accident given a variety of dependent variables including weather conditions, light conditions, road conditions etc.

➤ Interested parties

- Clearly, the police force/government agencies would be very keen on these predictions so that they can educate the public and advise caution to travel (or not) and while travelling especially during risky conditions
- The public themselves will also be keen to know when they should avoid driving so as to reduce risks of getting into accidents that then lead to injuries and fatalities.



Data acquisition and cleaning



- Data sources
 - Example dataset (37 columns of which one is SEVERITYCODE – target variable)
- Data cleansing
 - Imbalanced data set - 136,485 data points with Severity Code 1 and only 58,188 data points with Severity Code 2
 - Need to convert categorical variables to numerical values
 - Deal with missing values – either remove/replace
 - Look to add value with new column time of day from INCDTTM

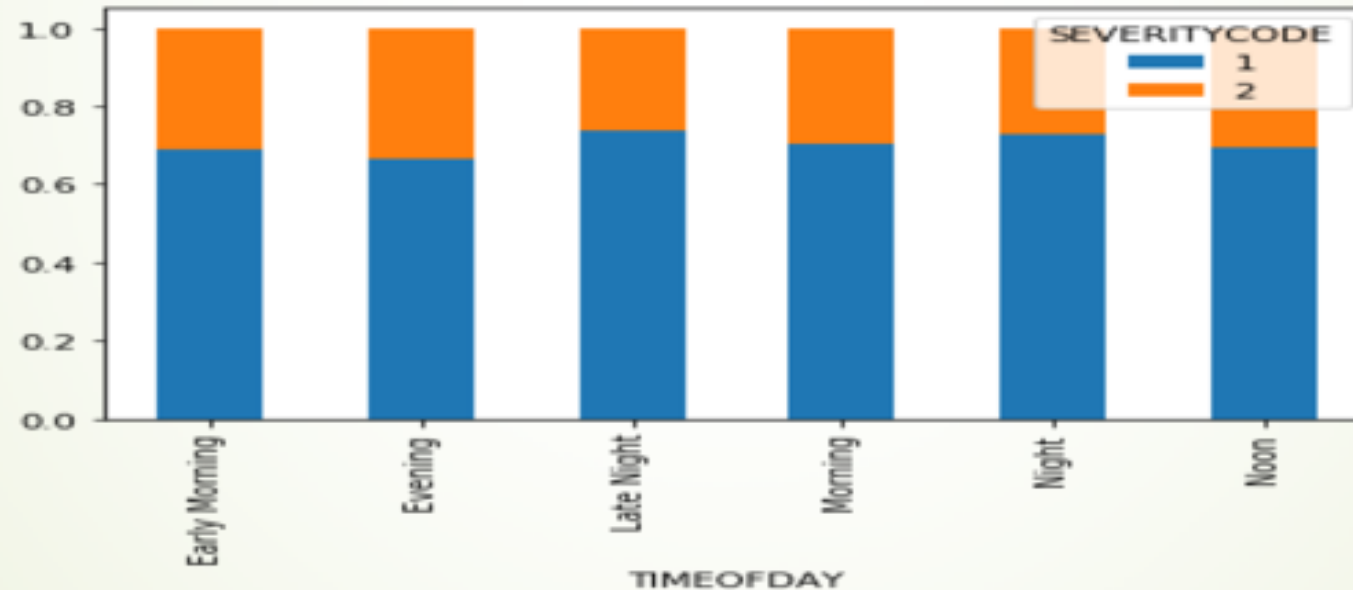
Data acquisition and cleaning – Final dataset used

Kept features	Dropped features	Added features
ADDRTYPE	X	TIMEOFDAY
COLLISIONTYPE	Y	
PERSONCOUNT	INCKEY	
PEDCOUNT	COLDETKEY	
PEDCYLCOUNT	REPORTNO	
VEHCOUNT	STATUS	
JUNCTIONTYPE	INTKEY	
SDOT_COLDESC	LOCATION	
INATTENTIONIND	EXCEPTRSNCODE	
UNDERINFL	EXCEPTRSNDESC	
WEATHER	SEVERITYCODE.1	
ROADCOND	INCDATE	
LIGHTCOND	INCDTTM	
PEDROWNOTGRNT	SDOTCOLNUM	
SPEEDING	SEGLANEKEY	
HITPARKEDCAR	CROSSWALKKEY	

Exploratory data analysis

➤ Relationship between time of day and accident severity

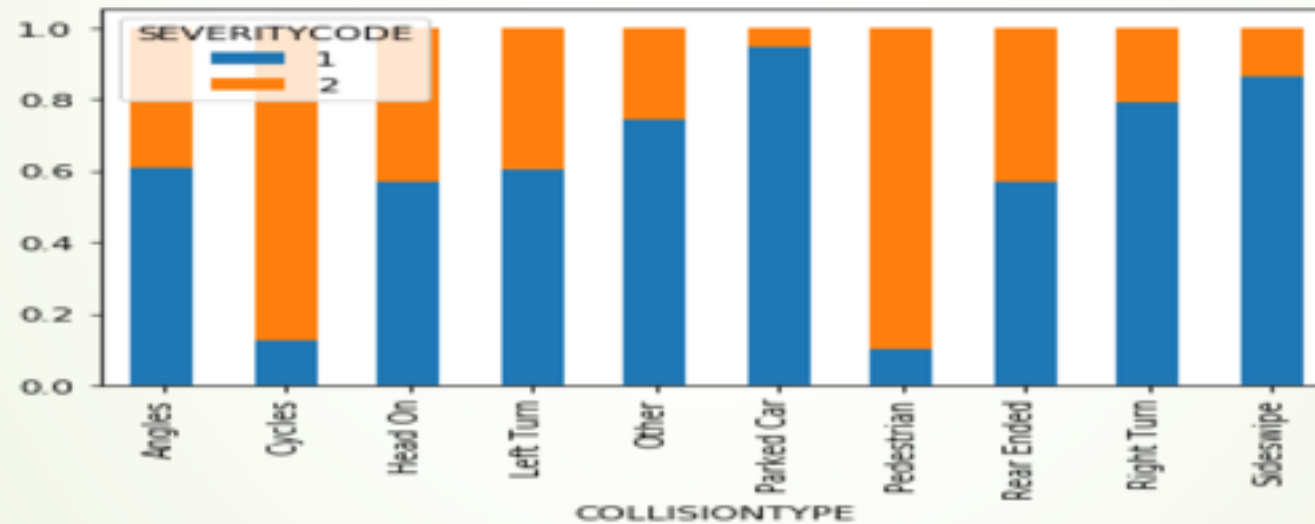
- Evening time slot (4pm to 8pm) has the highest percentage of accidents of severity code 2 (injuries)



Exploratory data analysis

► Relationship between COLLISIONTYPE and accident severity

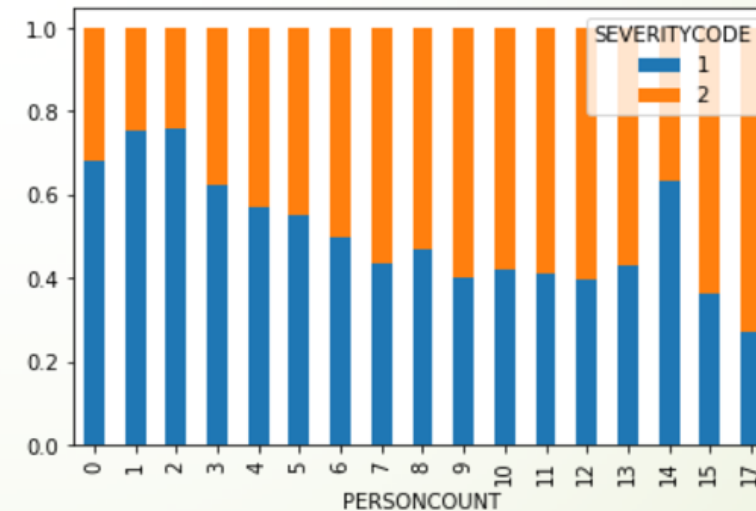
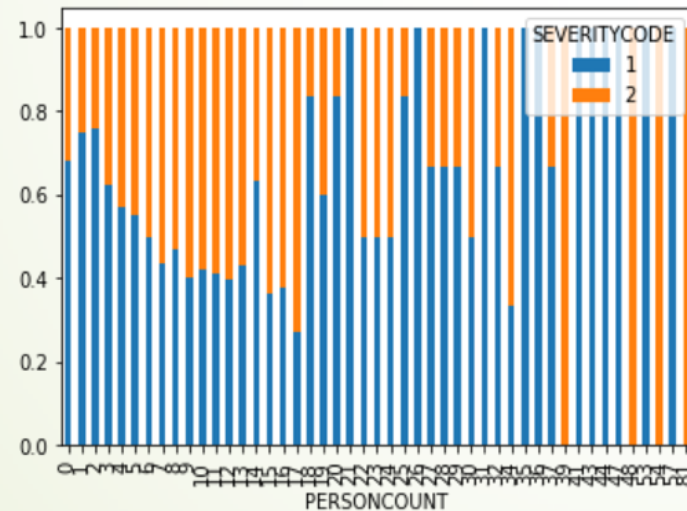
- Accidents which involve cyclists and pedestrians will result in higher severity as they are very exposed (not protected by a vehicle for example)



Exploratory data analysis

➤ Relationship between PERSONCOUNT and accident severity

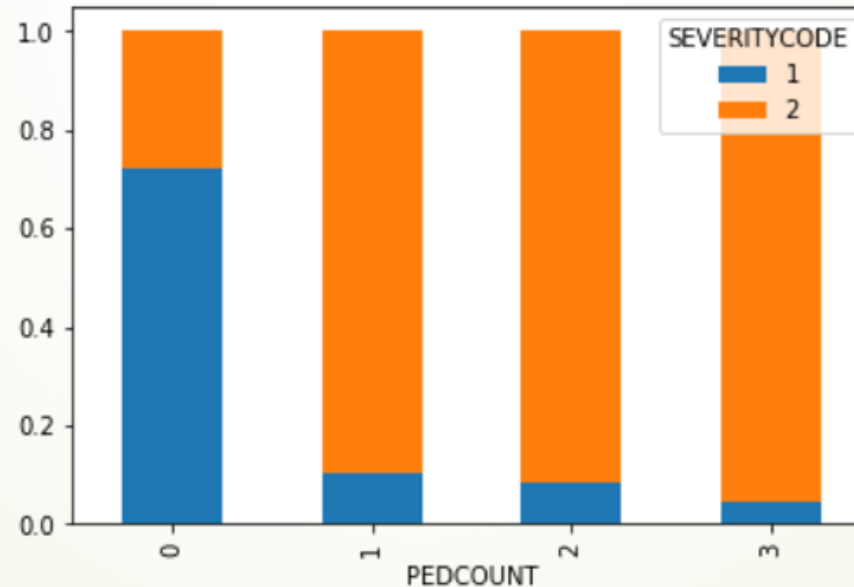
- The higher the person count, the higher the instances of an accident involving injuries
- Data below in chart (left) was cleansed further (see right) to remove counts with very small sample



Exploratory data analysis

➤ Relationship between PEDCOUNT and accident severity

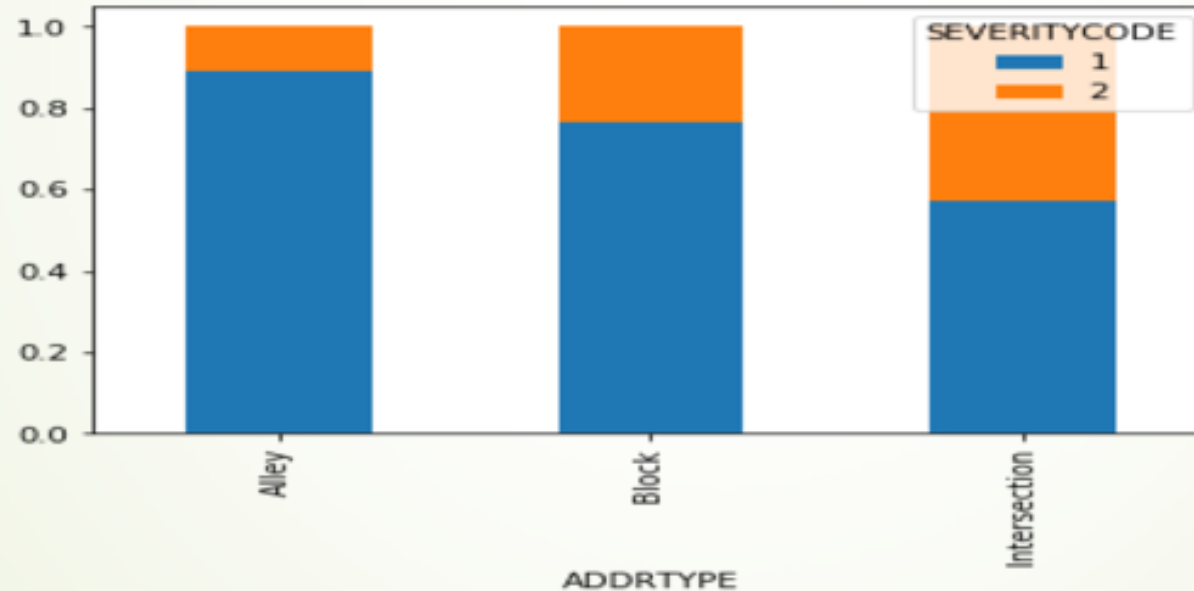
- Intuitively speaking, one would think that the higher the pedestrian count, the larger the severity of the accident and that is support by the chart below



Exploratory data analysis

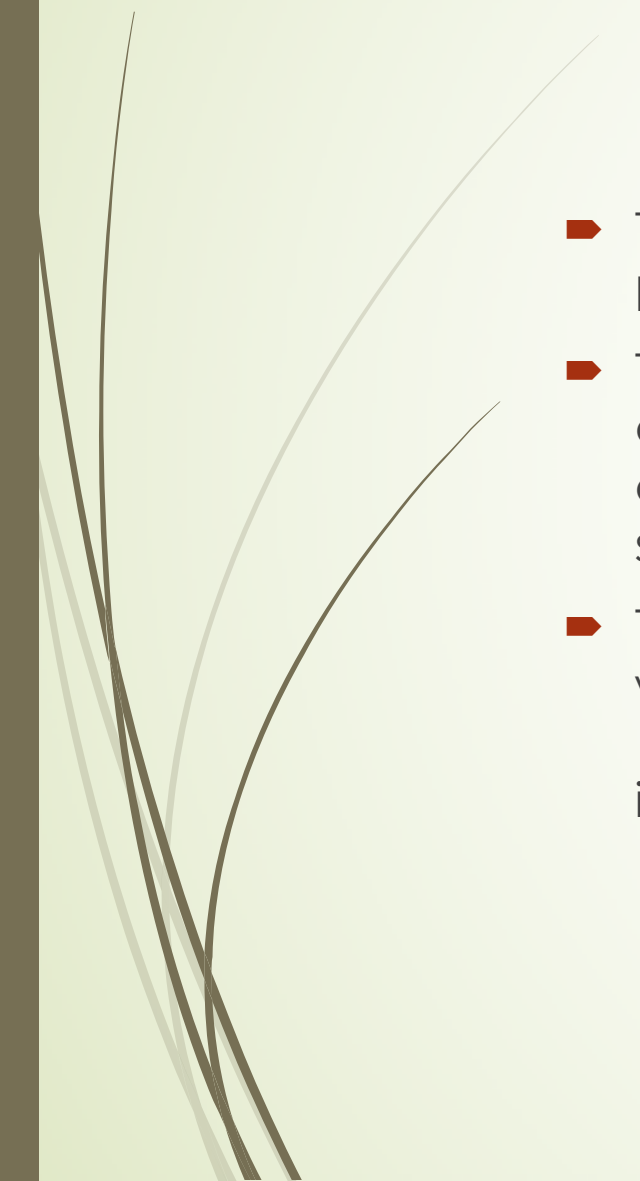
➤ Relationship between ADDRTYPE and accident severity

- Intuitively, accidents that happen at intersections will likely result in higher severity – this is supported by the chart below.





Predictive modelling – Decision tree

- This is in essence a classification problem where we are trying to assign predicted severity codes to accidents based on independent variables.
 - Therefore, a decision tree algorithm seems like the most suitable as the dataset is a sample of binary classifiers, and one can use the training part of the data set to build a decision tree and then use it to predict the severity of an accident.
 - The algorithm will use the most predictive feature to split the data set on. It works by selecting the best feature to decrease the impurity of the 'observations' in the leaves or to put it in other words, the feature that best increases the information gain (lowers entropy) after the split.
- 

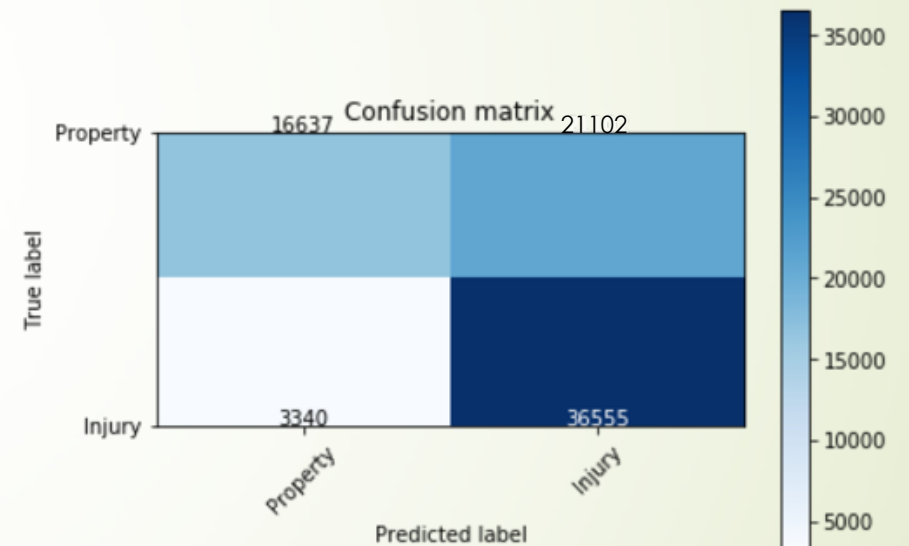


Decision Tree results

- ▶ The first feature used to separate the test data was COLLISIONTYPE-PARKED CARS. This is not a surprise as from the above observation, the separation of severity based on the parked car collision type is the largest and reduces the entropy the most.
- ▶ The second depth of features being used were PEDCOUNT and PERSONCOUNT – also not a surprise as both reduced entropy the most and resulted in the highest information gain
- ▶ Limitations of the decision tree model is that there are too many variables in this case such that having a max depth of 5 was insufficient to lower the entropy of the model to achieve a higher accuracy but going beyond 5, I was unable to generate a plot that was visible. Also there could be instances of overfitting

Decision Tree accuracy evaluation

- Using the decision tree model gave me an accuracy of the following:
 - Max depth = 100
 - Jaccard score: 0.7416080583249607
 - F1 score: 0.7374607245231716
 - Max depth = 5
 - Jaccard score: 0.68516371692815
 - F1 score: 0.6653808163743623





Conclusion

- In this assignment, I studied the relationship between severity of accidents and the prevailing conditions when the accident happened (time of day, type of accident, driver influence, weather, road, light conditions). I identified collision type and the person as well as pedestrian counts as among the key factors in determining the severity of an accident.
-
- I built a decision tree classification models to predict the severity of an accident. These models can be very useful in helping the police/government in a number of ways. For example, the police upon notification of an accident can predict whether the accident is likely to just be property damage or involve injuries and could react faster by calling an ambulance to stand by if it is predicted to be the latter.
-
- Of course, this model is by no means perfect will require inputs from experts and will have to involve further iterations.