A description of the problem and a discussion of the background. (**15 marks**)

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. Road accidents happen everyday and there may be certain common factors that contribute to an accident that results in injuries beyond just property damage.

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.

A description of the data and how it will be used to solve the problem. (**15 marks)**

Data acquisition and cleaning

Data sources
Data utilised is what is provided in the Example dataset. Data given include 37 columns of which one column is the target/dependent (i.e. SEVERITYCODE).

Data cleansing

- The dataset is slightly imbalanced, with 136,485 data points with Severity Code 1 and only 58,188 data points with Severity Code 2. Therefore, I use the methodology of upsampling to remove data imbalance to end up with an equal number of data points with Severity Codes 1 and 2 (i.e 136,485 data points each).
- Many of the variables provided are categorical – for example, weather conditions, road conditions, inattention etc. I therefore utilised replace function and one hot-encoding to convert these features into numerical representations.
- For some columns like SPEEDING, INATTENTIONIND and HITPARKEDCAR, there is only Y so I replaced the Nan with N – and converted them to binary 1s and 0s (1=Y, 0=N)
- After cleaning the data, whatever Nan values are left are dropped using the dropna function.
- There is a repeat column in that of SEVERITYCODE.1 which I dropped from my dataset. Also, in my opinion, certain columns (such as coordinates, object ID) were probably not very useful in predicting the severity of an accident – I opine that knowing the nature of the location of an accident (intersection, block) for example is more useful in determining the severity of an accident rather that the specific X,Y coordinates. I therefore dropped those columns as well.
- I added one new column – Time of Day – as time of day may be important too as visibility at night is lower than in the day and people may be more tired as well or peak periods for that matter will be important too as traffic is heavier.
- I experimented with Day of Week but it did not show any discerning properties.

Final feature selection

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	

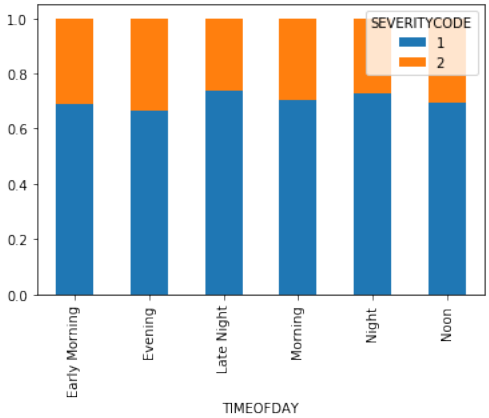
Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

Exploratory data analysis

Severity code 1 = Property damage
Severity code 2 = Injury

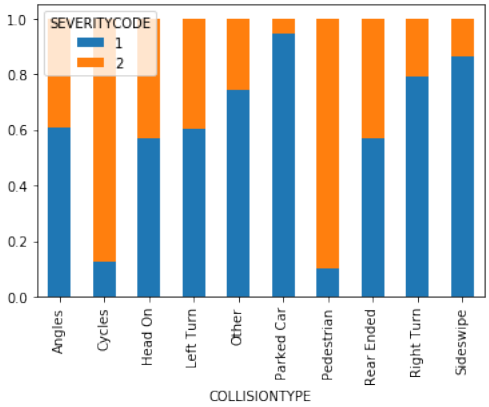
Relationship between time of day and accident severity

- Going into the analysis, I had thought that time of day would be an important predictor of accident severity where my initial hypothesis was that evening/late nights/peak periods may result in accidents that are more severe (i.e code 2 with injuries).
- Marginally, the chart below confirms that the evening time slot (4pm to 8pm) has the highest percentage of accidents of severity code 2



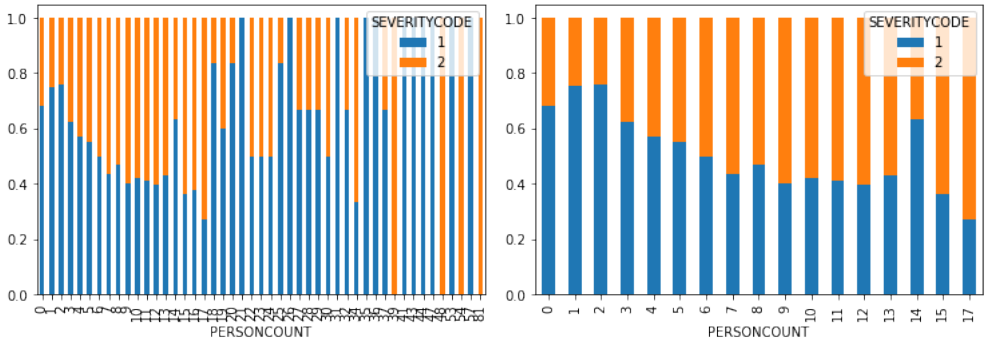
Relationship between COLLISIONTYPE and accident severity

- Intuitively, accidents which involve cyclists and pedestrians will result in higher severity as they are very exposed (not protected by a vehicle for example).
- This hypothesis is supposed by the chart below



Relationship between PERSONCOUNT and accident severity

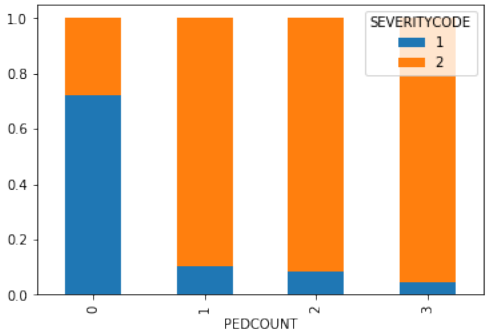
- Intuitively speaking, one would think that the higher the person count, the larger the severity of the accident and that is support by the chart below broadly where from person counts 0-17 we had generally an increasing percentage of accidents that were code 2 in severity.



- I note however that in some cases, there is 100% severity code 1 or 2 for person counts with higher values and that is because the number of observations with these values is very limited – one observation in some cases and not a large enough sample size and therefore probably inaccurate
- Therefore I dropped the PERSONCOUNT observations where the value_counts of the PERSONCOUNT is less than 10.
- Similarly, for SDOT_COLDESC, I dropped the observations where the value_counts were less than 100.

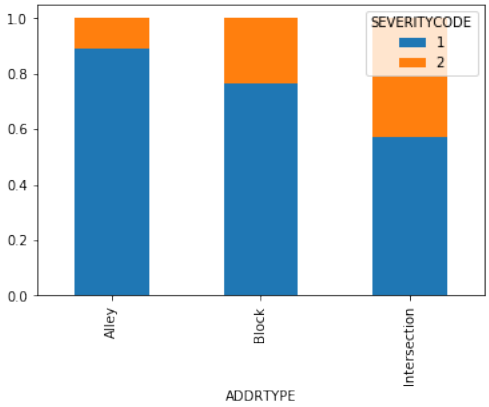
Relationship between PEDCOUNT and accident severity

- 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
- However, I note that PEDCOUNT of 4-6 have only very limited observations (less than 10 -> therefore those observations are removed as they could skew the results)



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

- 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
- 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

Conclusions

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.