Predicting Car Accident Severity based on Environmental Conditions

Benefits of reducing the frequency of severe accidents

- Major cities face numerous traffic accidents each year, with a range of severity from minor vehicle damage to multiple fatalities.
- Accident severity may be related to environmental conditions.
- Prediction of accident severity allows for:
    - Raising driver awareness of dangerous conditions to encourage safer driving
    - Allowing emergency responders to plan ahead when severe accidents are more likely
    - Reducing the the toll of accidents on numan life, municipal resources and economic damage

Dataset on traffic severity and other conditions

- Data set includes historical accidents in the Seattle area from 2004 onwards with data on the following:
  - Accident severity
  - Date and time of accident
  - Number of people/pedestrians/cyclists/vehicles involved and resulting injuries/serious injuries/fatalities
  - Presence of speeding/inattention/drugs/alcohol
  - Descriptive characteristics of the accident
  - Environmental conditions at the time of the accident including:
    - Weather
    - Road conditions
    - Lighting conditions

Data cleaning and preprocessing

- First, the dataset was narrowed down to just environmental variables, under the assumption that other variables are largely related to the occurence of the accident itself or human actions that are not relevant to the question at hand. Chosen variables are:
    - Accident Severity (target variable)
    - Weather
    - Road Conditions
    - Lighting Conditions
- Independent variables were categorical and converted to integer values to allow for modeling, nominally ranked by potential impact on driver ability
- Variables were then normalized and divided into a training and test set

Models tested

- Clustering models were chosen as the data was both labeled and the target variable (accident severity) possesses two distinct values in the available dataset (1 for less serious and 2 for more series
- Models chosen were:
  - K Nearest Neighbor (KNN)
  - Decision Tree
  - Logistic Regression
- For each model, different conditions were tested to find the optimal setup, arriving at the following:
  - KNN: 8 nearest neighbors
  - Decision Tree: depth = 3
  - Logistic Regression: all methods returned the same accuracy

Models performance

- Most models showed similar performance as seen below, with Decision Tree and Logistic Regression showing the best performance.
- Suggested model of choice is Logistic Regression as it provides the probability that an accident will be severe given a set of environmental conditions rather than a discrete result
- Model performance:

	Algorithm	Jaccard	F1-score	Log Loss
0	KNN	0.674544	0.546306	NA
1	Decision Tree	0.675558	0.544748	NA
2	Logistic Regression	0.675558	0.544748	0.62928

Conclusion and future direction

- It was shown that it is possible to provide a reasonably accurate prediction as to the severity of accidents based on environmental conditions
  - This allows for the initial goals of potentially being able to forewarn drivers of more dangerous conditions to increase awareness and allow for better planning of emergency services and reductions in cost in human life and economic damage
- However, there is still much room for improvement with an accuracy of under 70%. Potential ideals include:
  - More accurate, non-integer scaling of included environmental conditions based on impact on driver ability
  - Other factors such as time of day/time of year and examining the data at a more granular geographic split than the Seattle area as a whole