Introduction and Business Problem

Accidents occur often and the severity of the accidents typically requires a different response from authorities and other road users. Knowing the severity of a car accident can be incredibly important in helping road users make an assessment of how to navigate the road system to avoid exacerbating the situation. This can help road user re-route to alternate routes, and avoid congesting road ways where emergency services might need to use to provide critical aid to those hurt in an accident and likely save more lives.

Additionally, the impact of a severe accident to the other motorists cannot be understated. Predicting severity of accidents can help other motorists avoid such areas which will typically be congested and thereby saving motorists an inordinate amount of time. This will make for a generally pleasant commute time. The goal of this project is to develop a supervised machine learning model that would help a road user to predict car accident severity with reasonably high accuracy.

Data being used for this project

The data used in the project will the be the "Data-collisions.csv" file provided by the capstone project. The file has 194,673 rows and 38 columns. The target column is the Severitycode which has two states. 1 for slight accident with property damage and 2 for a severe accident with resulting injury.

Several key columns such weather, road condition and light conditions are missing data that will need to be imputed or dropped as necessary. There are several columns that are meaningingless and will need to be dropped from the dataset such as the object column which is merely a sequence numbering of the rows. This will be dropped during data preparation for modeling. Other columns such as reportno that do not add value to the modeling will also be evaluated and possibly dropped from the modeling during data preparation stage.

Severitycode column is duplicated. Data is imbalanced with more data showing cases where severity is lower i.e property damage compared with data showing a higher severity of injury. modeling will need to account for this appropriate i.e potentially using downsampling techniques.

Our modeling will perform a univariate and bivariate analysis of the different inputs to ascertain if they have a meaningful influence in the prediction. For example, the location

where the accident occurred, the weather and road conditions will be evaluated. Time of accident will also be important to evaluate to see if there is any correlation with severity of accident.

Methodology

The following steps were used in framing the business problem, gathering data, and then developing models for the data.

- Defined business problem
- Gathered data
- Prepared data for modeling (imputation of missing values and dropping some unnecessary columns)
- Built a decision tree classifier model, a random forest and a logistic regression model
- Performed model evaluation
- Obtained results and formulated a conclusion

Results

Below is the confusion matrix developed from the decision tree classifier model

The scores from each of the three models developed were similar indicating no model yielded better results compared to the others.

	Technique	Score
0	Logistic Regression	0.752782
1	Decision tree	0.753810
2	Random forest	0.753913

Discussion and Conclusion

- Models developed all had similar accuracy scores.
- The top three important features in predicting accident severity include

    - Number of pedestrians involved in the collision
    - Number of bicycles involved in the collision
    - The total number of people involved in the collision
- The three features account for 78% of the variance in the model.
- Model precision score is 0.76 and model recall score is 0.25. Model has high precision but low recall implying model is identifying high severity cases well but it also misses some high severity cases.