Predicting Car Accident Severity

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 on how to navigate the road system to avoid exacerbating the situation. This can help road user re-route to alternate routes and avoid congesting roadways 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.
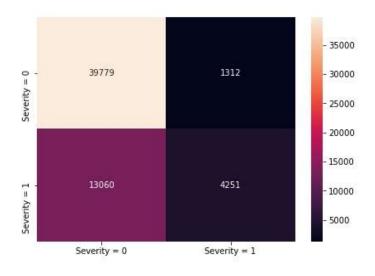
Data used in this project

- Utilized the "Data Collusions.csv" file provided for the project.

- Data file contains 194,673 rows and 38 columns.

- Target column is the SeverityCode which has two states:
  - 1 for slight accident
  - 2 for severe accident resulting in injury

- Several key columns such as weather, road conditions and light conditions have missing data that will be imputed

- Dropped several columns that have so much missing data and do not provide any meaningful contribution to the prediction modeling such as Incident date etc.

Methodology

- 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

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

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.