Introduction | Business Undertanding

In an effort to reduce the frequency of car collisions in a community, an algorithim must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

Data Understanding

Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

## Methodology

1. Data preparation and cleaning

Data cleaning procedure to make the dataset readable and suitable to the machine learning algorithms.

Dropping all the irrelevant variables and attributes

Out of the 37 attributes, I will not consider the features with over 40% of missing data, other unclear and irrelevant/noisy variables to our problem. I'll use COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND and UNDERINFL as attributes to classify SEVERITYCODE.

Dealing with missing values

As my chosen attributes have about 3% of missing data I'll just drop them. I'll still have a considerable amount of data.

Treating the categorical variables

In my case, all attributes are categorical. In this step, I will apply label encoding technique for all of them.

Train/Test split and data normalization

Now that I treated all my variables I'll separate my independent variables to dataset A and dependent variable 'SEVERITYCODE' to dataset B. After, I'll use this data to randomly pick samples and split in this ratio:

70% to train my model

30% to test my model Following the split I'll normalize all data to make sure my features are on a similar scale.

2. Classification: Modeling and Evaluation

The prepared dataset will be used to model 3 classification models.

Logistic Regression: Classifies data by estimating the probability of classes.

Decision Tree: Classifies by breaking down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

KNN: Classifies unseen data through the majority of its 'neighbours'. In this case we already know K=2 (2 classes of SEVERITY CODES). After obtaining each model's predictions we will evaluate their accuracy, precison, f1-score, log-loss and compare and discuss the results.

3. Discussion and Conclusion

After obtaining the results and evaluating them, in this section I will brief any observations noted based on the results. Finally, will conclude the results of this analysis.