**Introduction/Business Problem**

To reduce the frequency of bicycle-related collisions in a community, an algorithim will be developed to predict the severity of collisions involving cylists based on environmental factors, such as road condition, light condition, weather and accident location (intersection or block). This insight can serve to increase awareness of collisions in both cyclists and drivers under different conditions.

Because the predictor, SEVERITYCODE is either a 1 (property damage only) and 2 (involving injury), and three attributes have been selected in the model, the choice of machine learning algorithm will be Decision Tree and Logistic Regression. Performance between the 2 algorithms will be compareed.

**Data**

To perform the analysis, the following data will be needed:

1. To isolate the data on cycling-related collisions, ST\_COLCODE (collision code) = 45 (cyclist) is used as a filter on the dataset.
2. Only attributes ADDTYPE (address type), LIGHTCON (light condition), ROADCON (road condition) and WEATHER are kept in the dataset, along with the predictor SEVERITYCODE.

**Method**

After creating the subset of data, the number or records were reduced from 194673 to 4701. Records with NaN were dropped, resulting in 4687 records remaining.

Next, the number of unique values were counted for each attribute to understand their distribution. Severe over or under-representation of a value for an attribute would hinder predictive accuracy of the model. To achieve a more balanced and meaningful data reprsentation, records containining values in red were discarded and only values in bold were kept. Values that were discarded included “unknown” and “other” which provided little value for the prediction. Other discarded values which very low frequencies were also discarded.

ADDRTYPE

**Intersection 2651**

**Block 2029**

Alley 7

LIGHTCOND

**Daylight 3467**

**Dark - Street Lights On 794**

Dusk 188

Dawn 93

Unknown 87

Dark - No Street Lights 26

Dark - Street Lights Off 24

Other 1

WEATHER

**Clear 3160**

**Overcast 602**

**Raining 453**

Unknown 38

Other 2

Fog/Smog/Smoke 2

Snowing 2

Blowing Sand/Dirt 1

Sleet/Hail/Freezing Rain 1

ROADCOND

**Dry 3497**

**Wet 688**

Unknown 15

Ice 7

Standing Water 3

Snow/Slush 3

Other 1

Sand/Mud/Dirt 1

As a result, the following attributes and their remaining values used for prediction:

ADDRTYPE (2) – Intersection, block

LIGHTCOND (2) – Daylight, Dark – Street lights on

WEATHER (3) – Clear, overcast, raining

ROADCOND (2) – Dry, wet

This procedure further reduced the number of records from 4687 to 4185.

Imbalance of the predictor, SEVERITYCODE was also identified. There were 3702 records with SEVERITYCODE of 1, and 483 records with 2, which means there were 7.6 times more records with SEVREITYCODE of 1 than those with 2. To balance this dataset, the resample module from sklearn.util was used. In this case, upsampling was used to increase the numbre of records with SEVERITYCODE of 2 to 3702, thus bringing the total number of records to 7404. Finally, the entire dataframe is converted to string and reindexed.

For the Decision tree, entropy was used as the criterion and because there were only 4 attributes, the maximum tree depth was set as default.

For the logistic regression, large linear classification (liblinear) was used

**Result**

The F1-score and the Jaccard similarity score were used to evaluate the performance of the decision tree and Logistic Regression model

According to the F1-score, the decision tree predicted SEVERITYCODE of 1 and 2 53% and 57% of the time, respectively, with an average accuracy of 55%. Meanwhile, the logistic regression model predicted SEVERITYCODE of 1 and 2 57% and 52% of the time, respectively, with an average accuracy of 55%.

The Jaccard similaity score for the decision tree and Logistic regression were 0.5503 and 0.5483.

The decision tree:

**A screenshot of a cell phone

Description automatically generated**

**Discussion**

The predictive accuracy of both models were quite low at 55%. Changing the hypreparametr (c) fo the logistic regression and changing the maximum tree depth did not drastically improve the accuracy.

This suggests that there are other variables (causes) that result in the 2 severity types not that were not accounted by this model. In other words, the severity of accidents occurred mosty by chance given the environmental factors used in this model. In the original data set, driver charactersitics such as unattentiveness and driving under the influence were available. These attributes would be relevant in the outcome of an accident. However, for the purpose of this analysis, only environmental factors were of interest. Perhaps a more comprehensive model including more attributes would result in a higher predictive accuracy.

Reducing the number of low frequency values for the attributes led to a decision tree that is 5 levels deepe which makes it much easier to interpret.

**Conclusion**

The analysis demonstrated that light condition, road condition, weather and location of accident can only predict the severity of an accident 55% of the time. Human factors such as driver inattentiveness and driving under the influence would also alter the severity of an accident. The take home message of this analysis is that for cyclists and drivers to understand that accidents invovling cyclists are interactions of both their environment and driving behavior. One might not be able to avoid driving or cycling in averse environments all the time, but cyclists and drivers can always be more attentive on the road all the time.