Data

The collision dataset is collected by SDOT Traffic Management Division, Traffic Records Group in Seattle, which includes all collisions provided by SPD and recorded by Traffic Records from 2004/01/01 to 2020/05/20.

Data understanding: There are 194673 records of collisions and 38 attributes in the dataset. The **target variable** is SEVERITYCODE, which is **unbalanced** as 136485 records belong to category 1 (property damage) and 58188 records belong to category 2 (injury).

There are 5 variables selected to be explored and potentially used as input in modelling.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Description | # of missing values | Type |
| ADDRTYPE | Collision address type | 1926 | Categorical |
| JUNCTIONTYPE | Category of junction at which collision took place | 6329 | Categorical |
| WEATHER | A description of the weather conditions during the time of the collision | 5081 | Categorical |
| ROADCOND | The condition of the road during the collision | 5012 | Categorical |
| LIGHTCOND | The light conditions during the collision | 5170 | Categorical |

These variables will be further explored and analyzed regarding their distribution and correlation with the target variable, collision severity.

Data Preparation: Two to four variables out of 5 will be selected, cleaned and transformed if necessary, to build the predictive classification model.

Modeling: K-Nearest neighbors, decision trees, logistic regression and support vector machine will be trained based on the dataset, which will be divided randomly into training and testing datasets.

Evaluation: The project will select the best classifier based on their performance on the testing dataset. The metrics include Accuracy, Jaccard index and F1-score.

Deployment: The classifier will be used by the department of transportation and traffic to send more effective warnings of potential collision severity to drivers based on the conditions in different environments.

If there are further data released, the models will be reevaluated to see whether and how the prediction performance can be improved.