Data description

To evaluate the objectives of this project, a dataset from SDOT Traffic Management Divisions, Traffic Records Group in Seattle. This data has been provided from the organization's website through. This data is updated weekly and captures all types of collisions since 2004. A record of severity of the collision is provided alongside other 36 attributes of a given collision. The dataset has a record of 194673 collisions. For the purpose of this project, we will explore the location, weather condition, car speeding, light conditions, road condition, junction, number of people involved in the accident and the number of vehicles involved in the collision as the explanatory factors to predict severity of an accident. It would be expected however based on data cleaning, more features could be obtained through feature engineering. The dependent or target variable is severity of the collision which has been coded as follows; 1 – Low and 2 – High. The proportion of Low is 70.10% while that for High is 29.89%. This is an imbalanced dataset which will need to be balanced to ensure valid predictions.

Exploratory data analysis will be carried out to identify relationships between the various characteristics in the dataset with the target variable, severity of accident. Missing data will also be investigated and corrected. A set of predictions will also be identified to use in the prediction model. Since our data is labeled, supervised machine learning techniques will be used. The data will be split into training and test sets for training and validation. Since our target variable is binary, classification models will be evaluated to select the best classifier. The following classification models will be evaluated; K – nearest neighbor, Naïve Bayes, Random Forest, Logistic Regression, Gradient Boosting, XGBoost and Support Vector Machines. Various measures of accuracy such as Jaccard index, F1 score and log loss will be used to select the best classification model.