## Data

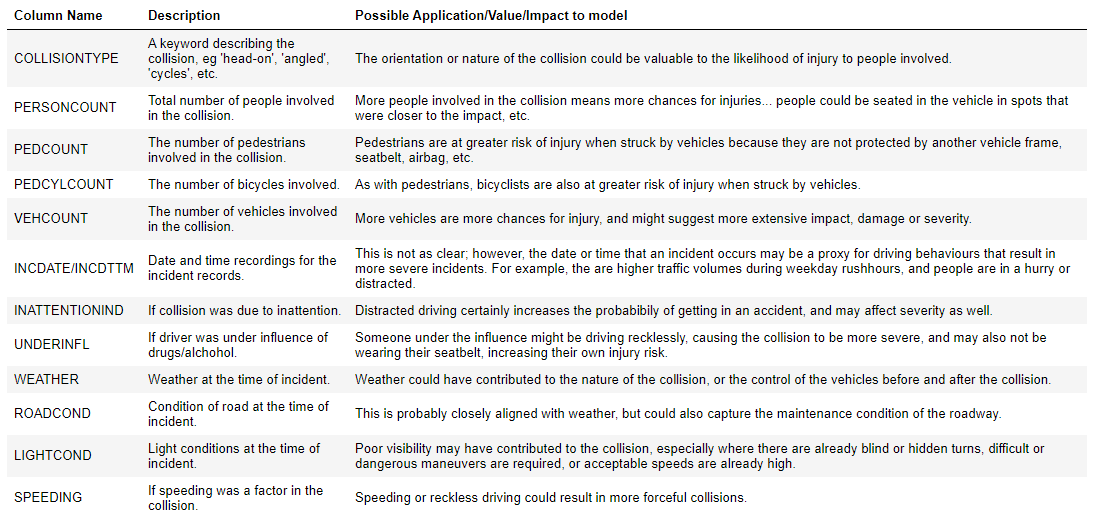
A model for predicting severity will be built using historic collision records collected by the Seattle Police Department and maintained by the SDOT Traffic Management Division. The dataset includes collision incidents from between 2004 and May of 2020. Records are assigned to two groups – ‘property damage only’ and ‘injury’ collisions – which makes it ideal for a supervised learning classification model. There are approximately 195,000 samples in the dataset; however, there are 2.35 ‘property’ collisions for every 1 ‘injury’ collision, so training subsets will need to be balanced in order to avoid skewing the model according to historic volumes.

  
*Figure 2 – count of collisions by classification, demonstrating the need to balance the data for training*

The dataset describes each collision using 36 different features, which provides a robust set of independent variables as possible parameters for training the model. These features identify:

* the date, time, position and location details of the collision;
* codes used by the state and DOT to categorize the collision;
* characteristics of the incident, such as type of collision and the number of people, pedestrians, bicycles and vehicles involved;
* environmental factors like weather and road condition; and
* the existence of driver behaviors including speeding, inattentiveness and the presence of drugs or alcohol

The table below lists the 12 features from the dataset that are possible candidates for training the model, along with high-level considerations regarding their potential value. Data exploration steps described in the Methodology section will demonstrate how these features were evaluated to determine possible correlation with collision severity and inclusion in the training features.

  
*Table 1 – candidate model training features from the SDOT collision dataset*