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.

```
Type 1 collisions (property damage) =  136485
Type 2 collisions (with injury) =  58188
Ratio of Collisions with property damage only to those with injuries: 2.35
```

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.

Column Name	Description	Possible Application/Value/Impact to model
COLLISIONTYPE	A keyword describing the collision, eg 'head-on', 'angled', 'cycles', etc.	The orientation or nature of the collision could be valuable to the likelihood of injury to people involved.
PERSONCOUNT	Total number of people involved in the collision.	More people involved in the collision means more chances for injuries... people could be seated in the vehicle in spots that were closer to the impact, etc.
PEDCOUNT	The number of pedestrians involved in the collision.	Pedestrians are at greater risk of injury when struck by vehicles because they are not protected by another vehicle frame, seatbelt, airbag, etc.
PEDCYLCOUNT	The number of bicycles involved.	As with pedestrians, bicyclists are also at greater risk of injury when struck by vehicles.
VEHCOUNT	The number of vehicles involved in the collision.	More vehicles are more chances for injury, and might suggest more extensive impact, damage or severity.
INCDATE/INCDTTM	Date and time recordings for the incident records.	This is not as clear; however, the date or time that an incident occurs may be a proxy for driving behaviours that result in more severe incidents. For example, the are higher traffic volumes during weekday rushhours, and people are in a hurry or distracted.
INATTENTIONIND	If collision was due to inattention.	Distracted driving certainly increases the probabibily of getting in an accident, and may affect severity as well.
UNDERINFL	If driver was under influence of drugs/alchohol.	Someone under the influence might be driving recklessly, causing the collision to be more severe, and may also not be wearing their seatbelt, increasing their own injury risk.
WEATHER	Weather at the time of incident.	Weather could have contributed to the nature of the collision, or the control of the vehicles before and after the collision.
ROADCOND	Condition of road at the time of incident.	This is probably closely aligned with weather, but could also capture the maintenance condition of the roadway.
LIGHTCOND	Light conditions at the time of incident.	Poor visibility may have contributed to the collision, especially where there are already blind or hidden turns, difficult or dangerous maneuvers are required, or acceptable speeds are already high.
SPEEDING	If speeding was a factor in the collision.	Speeding or reckless driving could result in more forceful collisions.

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