2. Data

The dataset of car accidents in the Seattle city (from 2004 to present) can be found from the below link. All collisions have been provided by the Seattle Police Department and recorded by Traffic Records.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The dependent variable is the accident severity in terms of 37 independent variables such as weather, road, and light conditions during the time of the collision. The target variable (SEVERITYCODE) is an integer number whose values can be either 0 (unknown), 1 (property damage), 2 (injury), 2b (serious injury), or 3(fatality).

2.1 Feature Selection

In the dataset, there are 37 independent features some of which might not be helpful to be used for the machine learning model. In this work, five features have been selected to predict the car accident severity. The first feature is “ADDRTYPE” that determines the address type of the collision (“Alley”, “Block”, and “Intersection”). The second feature is “COLLISIONTYPE” that is the collision type (e.g., sideswipe, parked car). The third feature is “WEATHER” that describes the weather condition during the time of the collision. The fourth feature is “ROADCOND” that is the condition of the road during the collision. The fifth feature is “LIGHTCOND” which describes the light conditions during the collision. The type of all independent features is *object* (i.e., text) while the type of the target variable is *int64*.

2.2 Data Preprocessing

Before using the data to train the model, the data should be prepared. First, the unnecessary columns should be dropped from the dataset. Next, we need to check whether or not the dataset is balanced. As can be seen from Figure 1, there are 136,485 samples with property damage while there are only 58,188 samples with injury. Thus, the dataset is unbalanced which means we should balance it to prevent training a biased model.

```
In [4]: df['SEVERITYCODE'].value_counts()

Out[4]: 1    136485
        2     58188
        Name: SEVERITYCODE, dtype: int64
```

Figure 1. unbalanced dataset

To balance the dataset, the samples with property damage (i.e., the majority samples) are down-sampled so that their number of samples is reduced from 136,485 to 58,188. As a result, the number of samples in both groups is equal that means the dataset is balanced.

Since the type of the independent features is *object*, we need to convert their type to numerical data types (i.e., *int64*). After doing these preprocessing steps, the data is ready to be used to train the model.