1. Data Acquisition and Cleaning

2.1 Data Source

Data used in this study is the collisions dataset provided by the Traffic Records Group, Seattle Department of Transportation (SDOT). The dataset includes all collisions provided by Seattle Police Department from year 2004 to May 2020.

2.2 Data Cleaning

It was fortunate that the data was in a nicely structured dataset. However, it is not without problems.

Before checking for missing values, some features deemed not necessary for predicting severity of traffic accident were dropped. They include unique identifier provided by SDOT, unique identifier of the incidents, unique identifier of the collisions, report number and other identifiers. As all records in this dataset are in Seattle, the coordinates feature was also dropped. Although the exact location where the accident occurred (intersection key, crosswalk key, lane segment) would be a great feature, there were too many unique cases under this feature, and hence was also dropped from the study.

Date and time variables were also provided in the dataset, however, while the dates are all complete, the time feature is not. It is generally perceived that time of the day could be an important factor in predicting traffic accidents and hence its severity, there are several other features provided in the dataset that imply the conditions at that moment, such as light conditions, road conditions and weather.

Several features provided in the dataset also points to the same thing. They include a code given to the collision by SDOT and the description of the code. The description is dropped from the study. There is also a code given to the collision by the state provided together with another variable of its description. The description is also dropped.

In terms of missing values, the variables for weather, road conditions and light conditions have significant numbers of missing values. We consider these as crucial predictors of traffic accident severity and hence decided to delete records with missing values for these 3 variables, instead of replacing them.

There were three indicator variables with missing values, i.e. whether or not the pedestrian right of way was not granted, Whether or not speeding was a factor in the collision, and whether or not collision was due to inattention. These variables only have entries for true cases and the missing values were replaced with false indicators.

Another indicator variable showing whether or not a driver involved was under the influence of drugs or alcohol has 4 unique entries, namely ‘Y’, ‘N’, ‘1’ and ‘0’. ‘Y’ and ‘1’ were grouped as positive cases while ‘N’ and ‘0’ were grouped as negative cases.

2.3 Feature Selection

The correlation among variables were studied and there were no 2 variables found to be highly correlated (Pearson correlation >0.9) and hence no features were dropped based on this method.

The final dataset contains 180,214 records with the following 17 predictors.

Feature	Description
ADDRTYPE	Collision address type: • Alley

	• Block • Intersection
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
INATTENTIONIND	Whether or not collision was due to inattention
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted
SPEEDING	Whether or not speeding was a factor in the collision
ST_COLCODE	A code provided by the state that describes the collision
HITPARKEDCAR	Whether or not the collision involved hitting a parked car

Among the 180,214 cases, 56,390 are considered severe cases while 123,824 are considered non-severe cases.