Methodology

Data

The source of data:

The data is a shared collection of all car accidents occurred in Seattle city from 2004 to Present. This includes all types of collisions provided by SPD (Seattle Police Department) and recorded by Traffic Records.

Description of data:

The raw data is composed of 37 attributes that describe 194673 different accidents. By looking at these attributes, they can be categorized as shown in table 1.

Table 1: Attributes of the Traffic Records provided by SPD.

Category #	Category name	Attribute's name
1	Identification	OBJECTID, INCKEY, COLDETKEY,
2	Location	SHAPE, ADDRTYPE, INTKEY, LOCATION, CROSSWALKKEY
3	Time	INCDATE, INCDTTM,
4	Accident's description	SEVERITYCODE, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, HITPARKEDCAR
5	driver's related causes of car accident,	INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING
6	Environment's related causes of car accident.	WEATHER, ROADCOND, LIGHTCOND
7	Unknown	EXCEPTRSNCODE, EXCEPTRSNDESC

As shown in table 1, three of these attributes are just to give identification number for the accident. Four attributes give the location and the address type. Two features inform when the accident occurred. Eighteen attributes describe the accident as how sever the accident was, the collision type, the number of vehicles involved, the number of people involved, the number of injuries and fatalities, and if the accident involved pedestrian and/or bicycles. Four attributes informed driver's related causes such as distracted driving (inattention), driving under influence of drugs or alcohol, speeding, and whether or not the pedestrian right of way was granted. Three attributes described the weather condition, the road condition, and the light visibility when the accident occurred. The last two attributes were miscellaneous with no clear meaning (unknown).

Attributes selection:

To select the important features that will be used in the analysis, it is important to clarify the question that needs to be answered. The question is that ***can the driver related causes and/or the environment's related causes of car accidents predict the severity of the accident that the driver could be involved into?*** Based on the previous question, it is clear that the target feature is the severity of the accident that should be predicted i.e (***SEVERITYCODE***). This target features is already coded in the data as "**One**" for the accidents resulting in property damage with no injury or "**two**" for the accidents resulting in injuries. Table 2 shows the selected independent features that will be used to predict the target feature. Four of these features are environment's related or ER and the other four are driver's related or DR. Based on the feature selection, all the other features were dropped from the data.

Table 2: The selected independent variables to predict the target variable.

Variable	Description
Target	
SEVERITYCODE	A code that corresponds to the severity of the collision 1- prop damage 2- injury
Independent	
Environment's related or ER	
ADDRTYPE	Collision address type: • Alley • Block • Intersection
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.
Driver's related or DR	
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.

SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)

Data cleaning

Multiple processes were done to the selected feature.

- **Removing the empty or non-meaningful data:** This was done by dropping all the accidents of empty data (i.e Nan) or without clear description such as values of “Unknown” or “Other”.
- **Categorical to numerical transformation**: Since all the values of the independent variables are of object type, the second process is to numerically coding all the text description as shown in table 3. The type of data was transformed from object to float

Table 3: The numerical coding of the data points of the independent variables

Independent variable	Values	Code
ADDRTYPE	Alley	**1.0**
	Block	**2.0**
	Intersection	**3.0**
WEATHER	Clear	**1.0**
	Partly Cloudy	**2.0**
	Overcast	**3.0**
	Blowing Sand/Dirt	**4.0**
	Severe Crosswind	**5.0**
	Fog/Smog/Smoke	**6.0**
	Raining	**7.0**
	Snowing	**8.0**
	Sleet/Hail/Freezing Rain	**9.0**
ROADCOND	Dry	**1.0**
	Sand/Mud/Dirt	**2.0**
	Wet	**3.0**
	Water	**4.0**
	Snow/Slush	**5.0**
	Ice	**6.0**
	Oil	**7.0**
LIGHTCOND	Daylight	**1.0**

	Dawn	**2.0**
	Dusk	**3.0**
	Dark - Street Lights On	**4.0**
	Dark - Street Lights Off	**5.0**
	Dark - No Street Lights	**6.0**
	Dark - Unknown Lighting	**7.0**
INATTENTIONIND	YES	**0.0**
	NO	**1.0**
UNDERINFL	YES	**0.0**
	NO	**1.0**
SPEEDING	YES	**0.0**
	NO	**1.0**
PEDROWNOTGRNT	YES	**0.0**
	NO	**1.0**

N.B. in the **UNDERINFL** variable, there were 4 types of values (“Y,N,1,0), so all “Y” values were replaced to 1.0 and all “N” values were replaced to 0.0.

- **Balancing the unbalanced data:** By running the histogram for the target variable to examine the overall distribution of values of the target variable across the data, it was obvious that the values of property damage with no injury (1) were almost double of the values of injury (2) as shown in figure 1. It is recommended to balance the data to avoid any biasness. To do that, the observations “1” in target variable were randomly shuffled and under sampled by randomly deleting some of the observations from that majority class, so the observations of the majority class can match the numbers with the minority class. Figure 2 shows the histogram of the target variable observations after balancing.

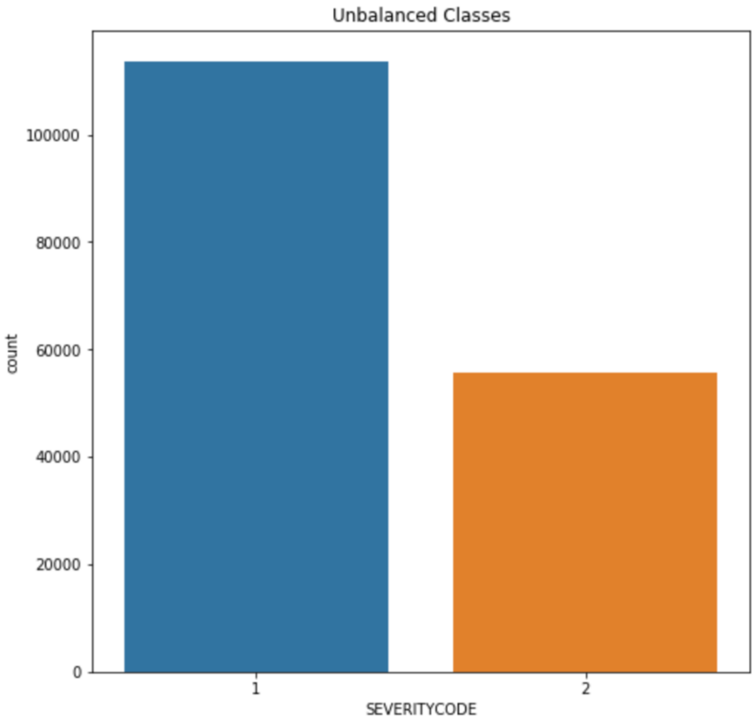

Figure 1: The histogram of the unbalanced distribution of the target variable observations (The severity of collision); 1: property damage with no injury, 2: injury.

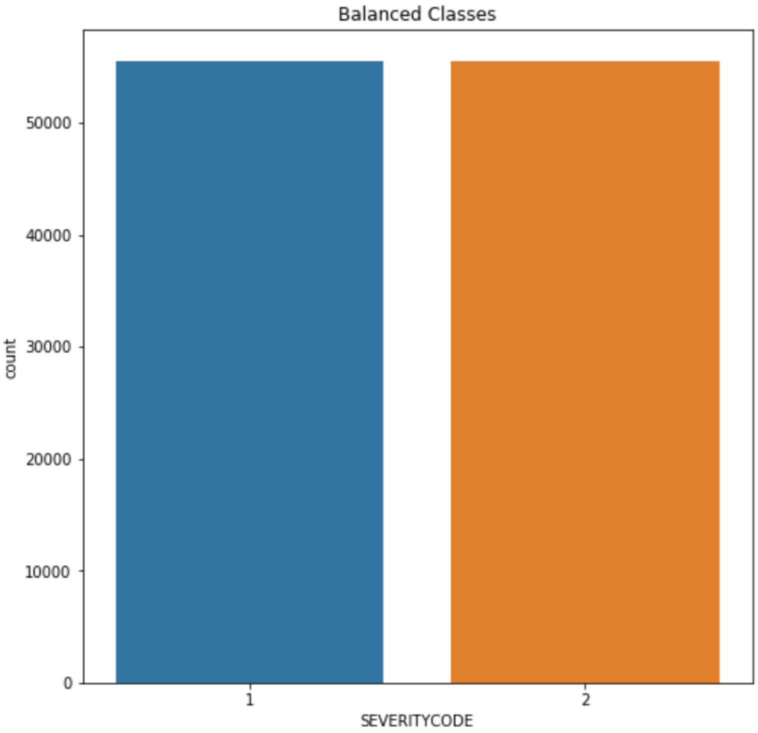

Figure 2: The histogram of the balanced distribution of the target variable observations (The severity of collision); 1: property damage with no injury, 2: injury.