CAR ACCIDENT SEVERITY PREDICTION

October 17, 2020

DATA

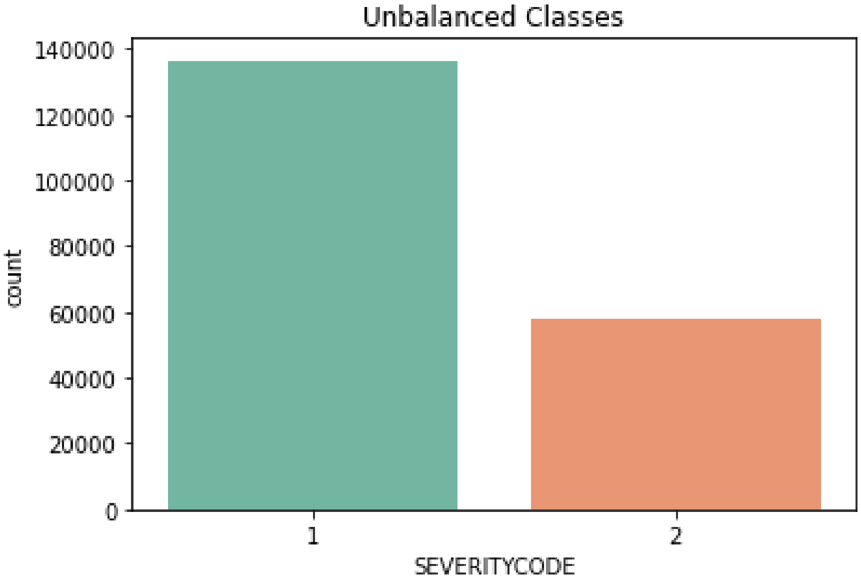
This project will work with data recorded by the Seattle DOF Traffic Management Division, Traffic Records group in the Data-Collisons.csv file which was downloaded from the Coursera website. The dataset has observations of traffic accidents from 2004 to the present day and is updated weekly. At the present time, there are approximately 194,673 observations in this data set with 38 different attributes. Non-essential attributes will be dropped from the dataset. A detailed description of the attributes can be found on the metadata.pdf in this project's GITHUB repository.

The target label for this project is SEVERITYCODE which has the following codes (according to the metadata.pdf):

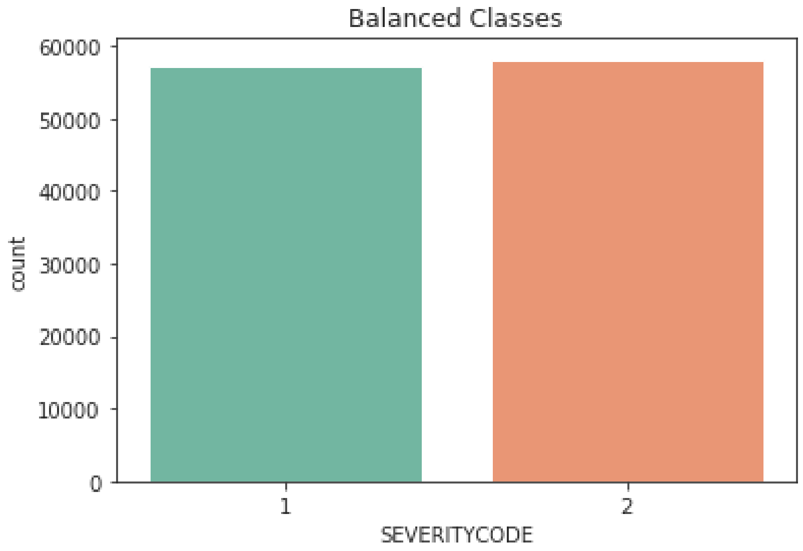
SEVERITYCODE

CODE	DESCRIPTION
3	Fatality
2b	Serious Injury
2	Injury
1	Property damage
0	Unknown

In the dataset, however, only severity codes 1 (Property Damage) and 2 (Injury) are used. There are 136,485 (70%) instances where severity code equals 1 and 58,188 (30%) instances where severity code equals 2. The dataset is imbalanced and will require resampling (undersampling) so that the dataset is balanced.



The graph below represents the dataset after resampling.



Upon initial exploratory analysis, the following is a list of important features in the dataset that will be used in the data model(s). Since many of the features contain classification information, they will have to be converted to numerical or binary format prior to running the model(s).

WEATHER – There are 11 different types of weather condition classifications in the dataset: Clear, Raining, Overcast, Unknown, Snowing, Other, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Dry	Clear, Partly Cloudy, Overcast
1	Wet	Raining, Snowing, Sleet/Hail/Freezing Rain
2	Unusual	Blowing Sand/Dirt, Severe Crosswind, Fog/Smog/Smoke
3	Other	Other, Unknown

ROADCOND – There are 9 different types of road condition classifications in the dataset: Dry, Wet, Unknow, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt, Oil. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Good	Dry
1	Fair	Wet, Sand/Mud/Dirt, Oil
2	Poor	Ice, Standing Water, Snow/Slush
3	Other	Other, Unknown

LIGHTCOND – There are 9 different types of light condition classifications in the dataset: Daylight, Dark – Street Lights On, Unknown, Dusk, Dawn, Dark – No Street

Lights, Dark – Street Lights Off, Other, Dark – Unknown Lighting. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Light	Daylight
1	Partial Light	Dusk, Dawn, Dark – Street Lights On
2	Dark	Dark – No Street Lights, Dark – Street Lights Off, Dark – Unknown Lighting
3	Other	Other, Unknown

SPEEDING – The speeding indicator denotes whether a driver was speeding by “Y”. This field will be converted to binary format. Fields containing a “Y” will be converted to a 1. Blanks or NaN will be converted to 0.

INATTENTIONIND – The inattention indicator denotes if the driver is inattentive with a Y. This field will be converted to binary format. Blanks or NaN will be converted to 0.

UNDERINFL – The under the influence indicator denotes if the driver is under the influence with either a Y or a 1. N, 0 and NaN denote not under the influence. This field will be converted to binary format.

Example of current values in balanced dataset:

```
N    60232
0    47463
Y     3340
1     2648
Name: UNDERINFL, dtype: int64
```

Once data is cleaned up, classification datatypes will be changed to integers for use in the model(s). The following is an example of the Feature set.

Target(Y)   Feature Set(X):

SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	INATTENTIONIND	UNDERINFL
2	2	2	2	1	0	0
1	0	0	0	0	1	0