**Introduction/ Business Problem**

The Washington State Department of Transportation Crash Data Portal provides crash information for accidents that occurred statewide. According to the 2019 data, there were 45,524 accidents on all roads. Of those:

* 235 were fatal crashes
* 973 were suspected of serious injury accidents
* 2,798 were suspected of minor injury accidents
* 9,412 were possible injury crashes
* 32,106 were no apparent injury collisions

Our motivation is to use the weather, location and road condition data provided in the dataset, made available by the Seattle Department Of Transportation Traffic Management Division, to arrive at a correlation to predict the severity of road accidents. This tool/data can then be made available to the public and the Seattle traffic authorities to possibly prevent/reduce severe or fatal accidents in the future by taking precautionary measures.

**Data Understanding**

We chose the unbalanced dataset provided by the Seattle Department Of Transportation Traffic Management Division with 194673 rows(accidents) and 37 columns(features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state(whether under influence or not), collision type. Hence we think its a good generalized dataset which will help us in creating an accurate predictive model.

The unbalance with respect to the severity code in the dataset is as follows.

|  |  |
| --- | --- |
| SEVERITY CODE | Count |
| 1 | 136485 |
| 2 | 58188 |

**Data Preprocessing**