Seattle Washington Car Accident Severity Case Study

T. Miglin

IBM Data Science Professional Certificate

Table of Contents

1. Introduction

1.1 Background Knowledge to the Problem

In this case study, data acquired about car accident severity in the city of Seattle, Washington is analyzed and modeled to better categorize each incident. Seattle is a landmark city in the United States and is home to about 3,433,000 people, with vehicle ownership being around 444,000 (Macrotrends.net). It was also reported that there is a direct positive correlation between population booms and number of vehicles. The given dataset only contains general location information, so while high density locations may seem pivotal to the study, they will not be included in the model.

1.2 The Problem

In 2018, 36,560 people were killed in motor vehicle accidents, and 6,227 of those deaths were pedestrians in traffic. These large death tolls span car to car collisions, car to environment collisions, pedestrians struck, drunk drivers, trains, cyclists, etc. The questions come to, how can this number begin to trend downwards, and what situations are leading to so many severe vehicle accidents?

See as the data based solution of this problem lies in categorizing accidents, the results of this study have the capability to be applied in many situations. DOT's and highway engineers could use this modeling to better design traffic systems as well as pedestrian safety. Emergency first responders local to the Seattle area may be able to better prepare equipment as well as necessary medical attention for scenes of accidents. Overall, the severity of these accidents can be reduced based on attending to an amount of major factors.

2. Data

2.1 Detailing the Dataset

The dataset given for the study is derived from the Colissions-All Years set from the SDOT Traffic Management Division, Traffic Records Group. Made up of 38 columns holding information on every type of collision from 2004 to present, including cars, bicycles, and pedestrians. Many of the columns in the table are either keys or codes for names of road structures, or categorical values regarding the environment of the accident. Also included in the metadata is a dictionary of collision codes with many codes. Some that are of similar incidents with different orientations or radial placements. For the sake of the model, most of these codes are not necessary as the goal is to purely know general severity.

2.2 Cleaning the Data

With the goals of predicting severity of accidents and finding ways to prevent severe accidents, it seems obvious that a classification model is necessary. The study will be using KNN, Decision Tree, and Logarithmic Regression to test and train the model. In order to build the feature set necessary, the target classifications need to be known. The target known to the study as Severity Code was encoded into the form of 0, which represents property damage only/ no injuries, and 1, which represents injuries sustained. The degree of injury will not come into question as the assumption is, any injury is a very bad injury.

2.3 The Feature Set

For the study there are 5 features included with the goal of classifying those features as a severity code. The 5 features are, LIGHTCOND, ROADCOND, WEATHER, SPEEDING, INATTENTIONIND, and UNDERINFL. Each of the features were encoded into integer values with 0 being assigned to the element with the lowest possibility of a severe accident. For example, in ROADCOND, 0 is assigned to dry and 2 is assigned to wet/icy. The yes or no features were assigned 1 and 0 respectively. One issue that arose is the difference in number of entries between INATTENTIONID and SPEEDING, and the rest of the features. In order for the study to be successful there needed to be a consistent value for unknown across the feature set. So all unknown values across the set become “Unknown” for consistency and better proportions. The feature set and encodings can be seen in the figure on the next page.

These features were selected since they can have major impacts on severity of an accident. Other factors can be seen as major in just causing an accident but these features are able to fit into both roles. Being inattentive or being under the influence can appear to be of similar degree since it would seem practical to say being inebriated makes you inattentive, but it does not translate both ways. All the other features directly affect driving ability and do not have to reflect on the driver's skill or attention, which can have a larger effect on changes in traffic patterns or fine limits.

Feature	**Encoding**
All Features	Unknown: Any NAN or other values
LIGHTCOND	0: Daylight 1: Dark (Street lights on) 2: Dark (No extra lighting)
ROADCOND	0: Dry 1: Bumpy, light friction 2: Slippery
WEATHER	0: Sunny, Clear 1: Partly Cloudy 2: Low visibility 3: Precipitation
SPEEDING	1: Yes 0: No
INATTENTIONIND	1: Yes 0: No
UNDERINFL	1: Yes 0: No

Figure 1: Encodement Table for Feature Set

3. Methodology

3.1 Data exploration analysis

Taking into account that the project's feature set and target are categorical variables with limited options for classification, the correlation between the values may view different from the actuality. In order to better understand the feature set selection, a few visualizations of difference in entry amounts will help. In the image below, the target variable is differentiated between property damage and physical injury. It is seen that property damage outweighs physical injury by more than double, and this study will not bear practical results without a balanced dataset. Here is where SMOTE is used to balance the distribution of the data so there is an unbiased classification model.

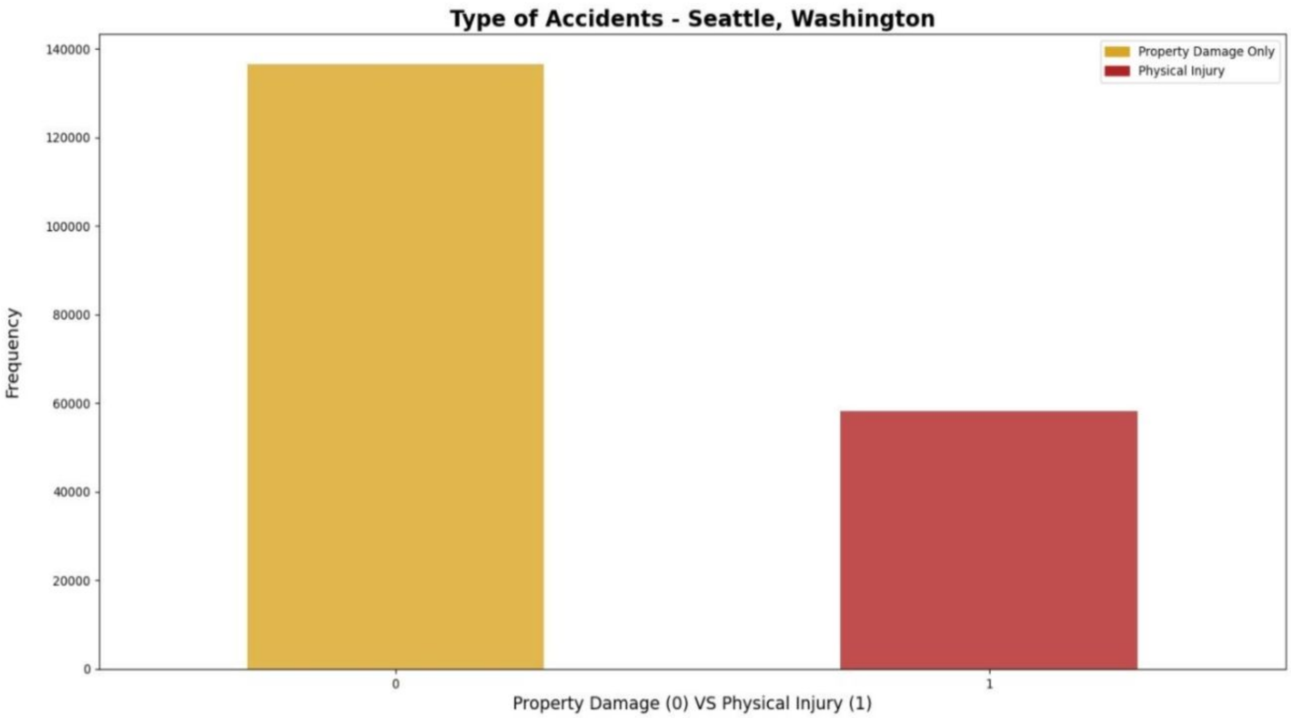

Figure 2: Target Variable Classification Bar Graph

Along with understanding the study's target variables, an expansion of knowledge on the feature set is also necessary. Thus, visualized below is the representation of all "non-zero", or least probable, values of the feature set. From this set it is clear that driving under the influence and over-speeding made up the minority of accident causes. The driving environment; weather, lighting, and road conditions, all correlate with each other, which makes sense of their similarly higher frequency.

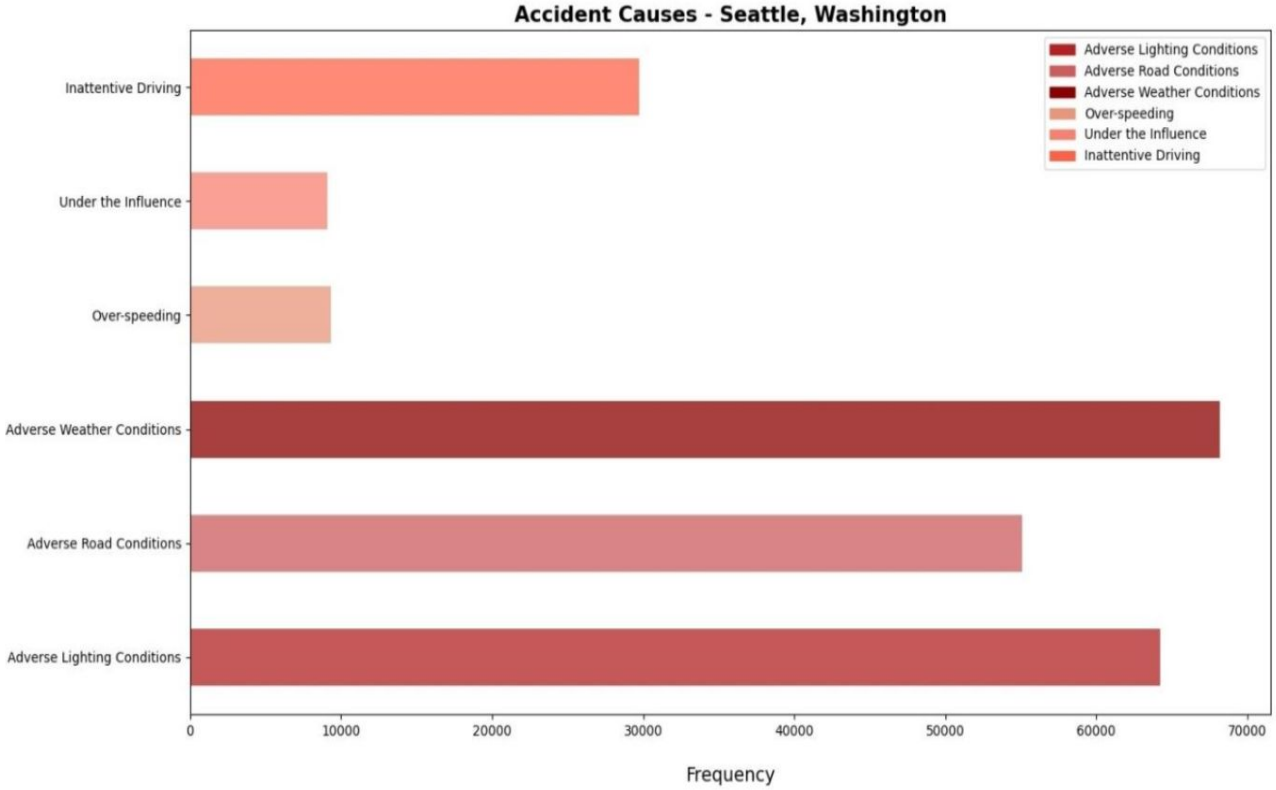

Figure 3: Frequency of Probable Accident Causes in Feature Set

3.2 Model Selection

The machine learning models used in this study are, KNN (k-Nearest Neighbor), Decision Tree, and Logistic Regression. KNN is an algorithm that relates all available cases to newly entered cases and classifies them based on a measure of closest distance in similarity score. Decision Tree analysis divides the data set into smaller subsets, all while a full decision tree is being built with decision nodes and leaf nodes. Logistic regression is modeled after a basic logistic function and builds on a binary dependent variable.

4. Results

4.1 k-Nearest Neighbor

Similar to the lab work from the machine learning with python course, the k-Nearest Neighbor classifier from the sci-kit learn library was used to train and test the model on the SMOTE balanced data set. Seen below is the graph depicting the best KNN value as well as the classification report which will be expanded upon later. Code for the graph is sourced from that specific lab.

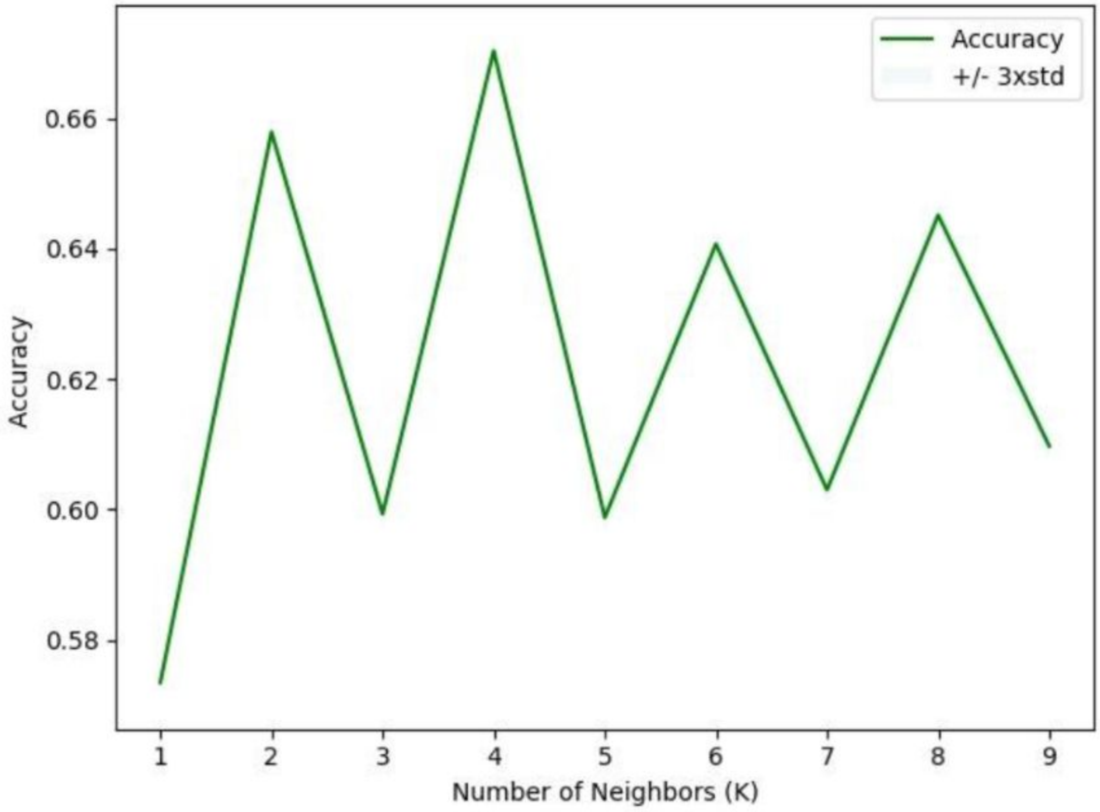

Figure 4: Best KNN value

	Precision	Recall	f1-score
0	0.93	0.7	0.8
1	0.08	0.32	0.1
Accuracy	0.67	N/A	N/A
Macro Avg	0.5	0.51	0.46
Weighted Avg	0.86	0.67	0.75

Figure 5: KNN Classification Report

4.2 Decision Tree

Similar to the lab work from the machine learning with python course, the Decision tree classifier from the sci-kit learn library was used to train and test the model on the SMOTE balanced data set. Seen below is the confusion matrix comparing the predicted severity label against the true label, as well as the classification report.

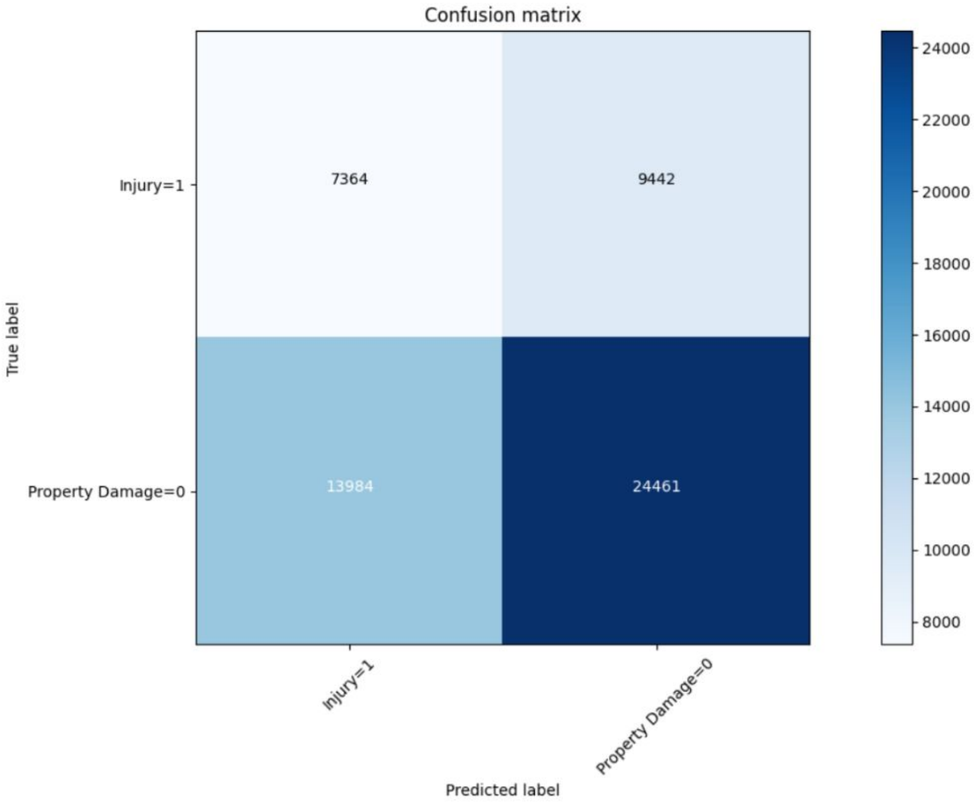

Figure 6: Decision Tree Confusion Matrix

	Precision	Recall	f1-score
0	0.64	0.72	0.68
1	0.44	0.34	0.39
Accuracy	0.58		
Macro Avg	0.54	0.53	0.53
Weighted Avg	0.56	0.58	0.56

Figure 7: Decision Tree Classification Report

4.3 Logistic Regression

Similar to the lab work from the machine learning with python course, the Logistic Regression classifier from the sci-kit learn library was used to train and test the model on the SMOTE balanced data set. Seen below is the confusion matrix comparing the predicted severity label against the true label, as well as the classification report.

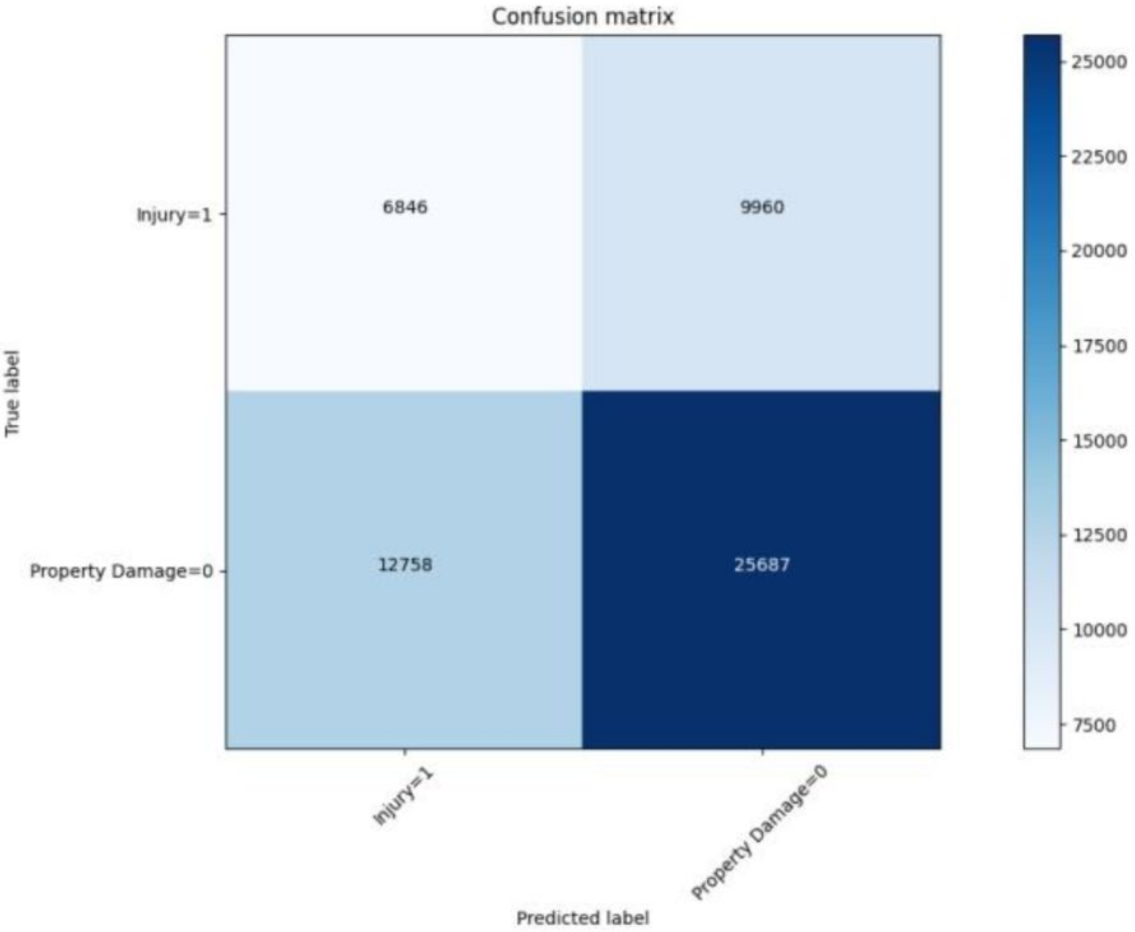

Figure 8: LR Confusion Matrix

	Precision	Recall	f1-score
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.60
Log Loss	0.68		

Figure 9: LR Classification Report

5. Discussion

The precision of a model is the percentage of results which are relevant, or better said, how many cases are similar disincluding outliers.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Working with Precision is recall, which is the fraction of the relevant documents that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Finding the harmonic mean of precision and recall gives what is called the f1-score, which is a measure of accuracy of a model.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

If precision and recall are perfect the f1-score will be shown as 1, which is also the highest score achievable. In the table below, the f1-score shown is the average of the binary elements of the target variable. k-nearest neighbors achieves the highest average f1-score meaning that it also achieves the highest precision and recall in hindsight. However, the average f1-score is deceiving since due to property damage outweighing physical injury in the data set. So, while KNN might have the highest average f1-score it is biased towards the precision and recall of the zero value.

6. Future Actions and Conclusion

6.1 Future Actions

After assessments of the data and the output of the models, there is some knowledge to be passed on to the stakeholders. The department of transportation for the city of Seattle can do a diligent evaluation of areas that may have a high concentration of accidents in, proper lighting, road conditioning, and effects of weather. These evaluations may result in the change of speed limits, traffic patterns, or addition of driving lanes. Additional signage can also be quite useful as alerting the driver that a certain area is a high risk crash area will allow them to take extra precautions.

6.2 Concentration of Accidents

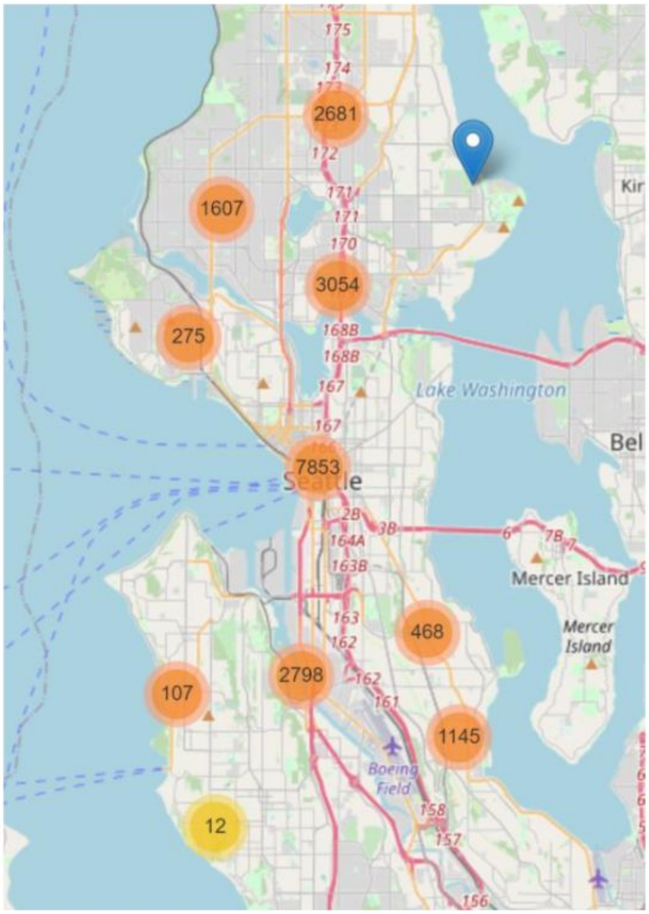

Figure 10: Folium Map of Seattle Accident Concentration

The higher concentration of accidents take place on higher density main roads of the city. Especially seen around the I-5 highway that runs through the center of the city. Most of those incidents also occur with bad weather, which will have a negative effect on road conditions. So it would behoove drivers to drive defensively and drive with caution under poor weather conditions on this main highway. Extra signage or inclusions of digital signage to give drivers alerts would be beneficial as a majority of accidents do not include an inattentive driver.

7. References

https://www.nhtsa.gov/

https://wsdot.gov/

https://ww.asirt.org/safe-travel/road-safety-facts/