PREDICTING ACCIDENT SEVERITY IN SEATTLE

1 SEATTLE’S VISSION ZERO PLAN

While Seattle is one of the safest cities in United States of America, there are still more than 10,000 crashes per year, resulting in average of 20 people losing their lives and over 150 people sustaining life-changing injures.

Despite the downward trend of fatalities and serious injuries, Seattle's Vision Zero Plan takes extra steps to ensure its citizens' safety. Indeed, traffic collisions are not accidents; They are preventable through smarter street design, targeted enforcement and etc. Hence, this paper is intended to support Seattle's ambition to bring down the accident fatality rate.

2 PROBLEM STATEMENT

In order to eliminate accident fatalities in Seattle, Seattle government need to

- To predict the severity of accidents
- To identify the primary factors affecting severity of accidents
- To reduce the impact of accident

Hence, an accident severity model will be built using machine learning algorithm, to predict the accident severity and identify the risk drivers causing fatalities. The results from the model may provide a good insight for Seattle government to reduce the severity of accident.

3 LITERATURE REVIEW

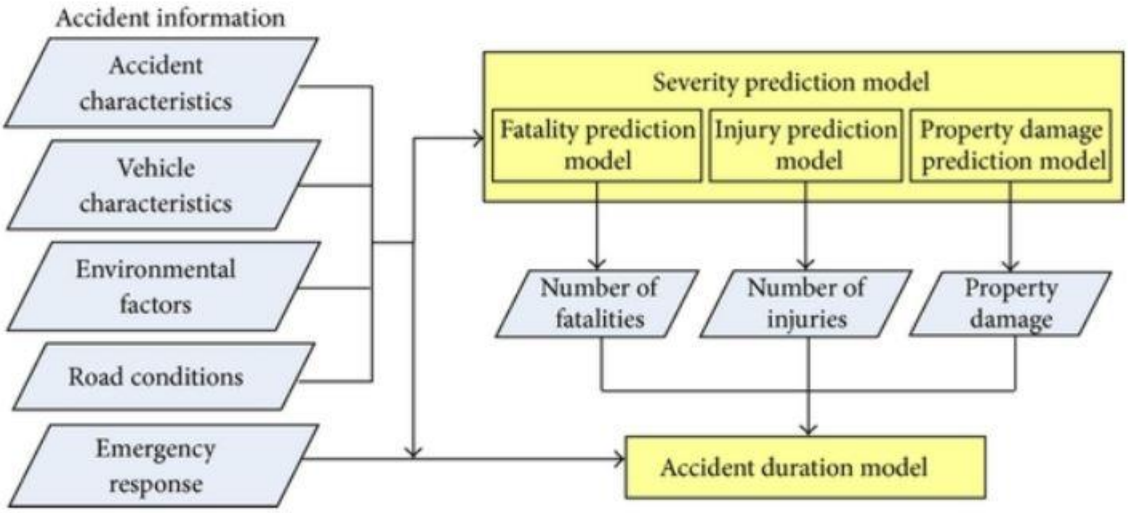


Figure 3.1: Accident Framework

Accident severity analysis has long been important topics for research and government. A study by Shankar, Mannering and Barfield (1996) had provided valuable evidence on the effects that environmental conditions, highway design, accident type, driver characteristics, and vehicle attributes have on accident severity. Furthermore, Praveena and Srinivas (2017) found that road features and driver's age were significant factors. Other accident studies also looked into severity types such as fatalities (Shibata and Fukuda, 1994) or crashes involving certain types of vehicles such as heavy trucks and combination vehicles (Alassar, 1988; Chirachavala, 1985). In a nut shell, the accident framework can be summarized in **Figure *3.1*: Accident Framework**Figure 3.1.

The accident framework above will serve as primary guide of this paper. Hence, this paper will consider the following dimensions as the potential causes of accident severity:

- Accident characteristics
- Vehicle characteristics
- Environmental factors
- Road conditions
- Driver characteristics

4 DATA PROCESSING AND EXPLORATION

To reserve the consistency of the review process in Capstone Project, this paper uses the Seattle's accident data that includes all collision types from year 2003 until now, to build the accident severity model in predicting the severity of an accident.

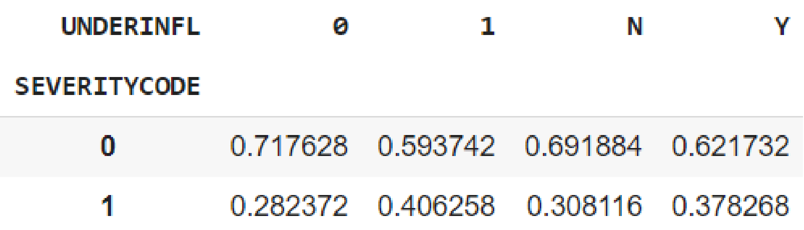
List of potential variables considered are as tabulated in table below.

Attribute	Description
ADDRTYPE	Collision address type
COLLISIONTYPE	Collision type
JUNCTIONTYPE	Category of junction at which collision took place
INATTENTIONIND	Whether or not collision was due to inattention
UNDERINFL	Whether or not a driver involved was under influence of drugs or alcohol
WEATHER	Weather conditions during the time of 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	Speeding
HITPARKEDCAR	Whether of not the collision involved hitting a parked car
INCDTTM	Use to derive weekend variables
SDOT_COLDESC	To derive the strucking vehicle

4.1 Data Cleansing

There are only two possible outcomes in the data, which is property damage (indicated by 1) and injury (indicated by 2). Hence, the target variables are redefined as injury and non-injury, which represented by 1 and 0 respectively.

Table 1: Inconsistent data entry in Under Influence variable



UNDERINFL	0	1	N	Y
SEVERITYCODE				
0	0.717628	0.593742	0.691884	0.621732
1	0.282372	0.406258	0.308116	0.378268

As shown in Table 1, Variable “Under influence” has inconsistent data entries. Since values of 0 and 1 have similar distribution as N and Y respectively, we believe that 0 and 1 may indicates N and Y respectively. Hence, normalization of variable “Under influence” into N and Y.

4.2 Feature Engineering

Number of vehicles on the road is expected to increase during the weekend due to family days and possible short trip of vacation. As the number of cars increases, the number of crashes increase, resulting in more severe accidents. Hence, variable “Weekend” is derived with 1 indicates weekend while 0 indicates weekdays.

Besides, the size of a strucking vehicle may determine the severity of an accidents. Thus, based on the state collision description, the variable “Strucking vehicle” are derived.

4.3 Treatment of Missing Data

Procedures to examine the missing data are as follows:

1) Examine whether any rows or columns contains all missing values
2) Examine whether any rows or columns pose high proportion of missing values, say 50%
3) For simplicity, complete case deletion approach may be considered given that the impact on the class distribution is negligible

Upon checking, all attributes have less than 1% missing values.

Table 2: Impact of complete deletion approach

Target variable	Before		After	
	Count	Proportion	Count	Proportion
Non-Injury	136,485	0.701099	126,270	0.690396
Injury	58,188	0.298901	56,625	0.309604
Total	194,673	1	182,895	1

Table 2 illustrates the impact of complete deletion approach. After deletion, the sample contains 93.95% of original observation. Hence, the sample bias is of negligible. Besides, the proportion of after deletion and proportion of before deletion are of similar. Since complete case deletion approach does not disrupt the class distribution. Thus, for simplicity, complete case deletion approach is used.

4.4 Data Exploratory Analysis

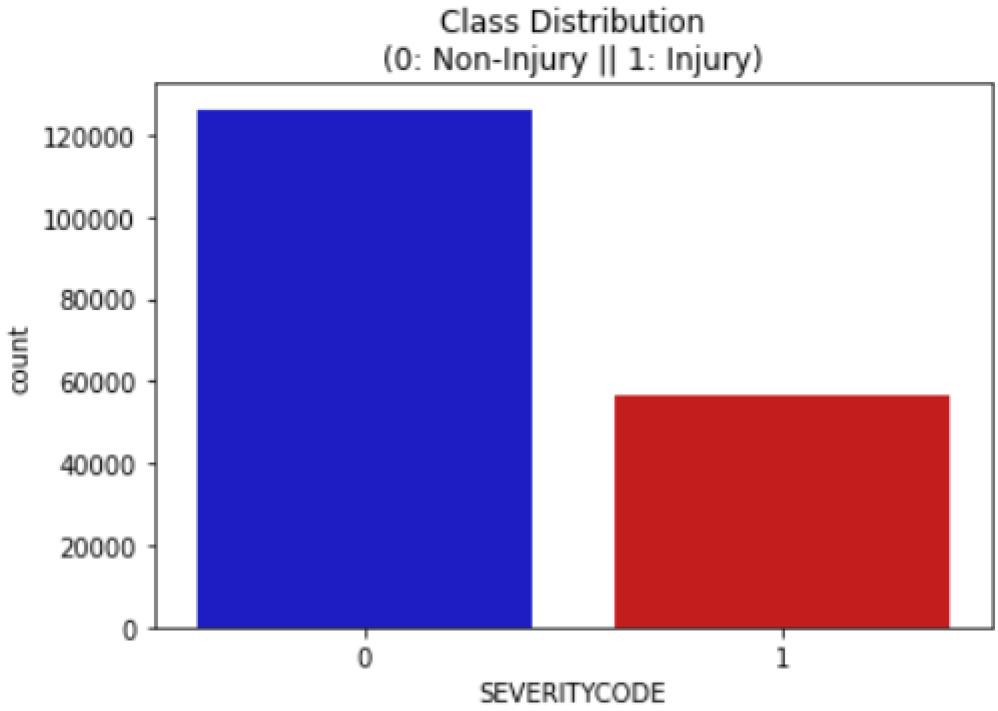
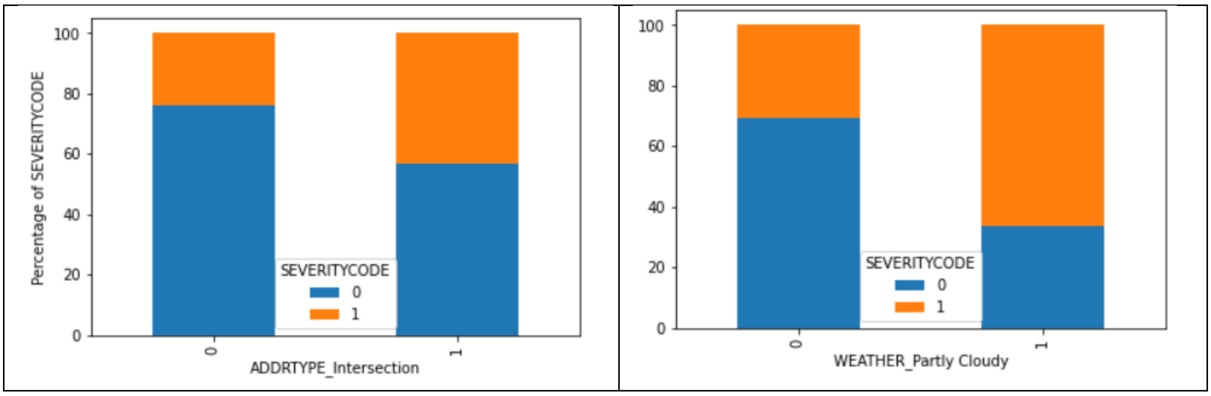
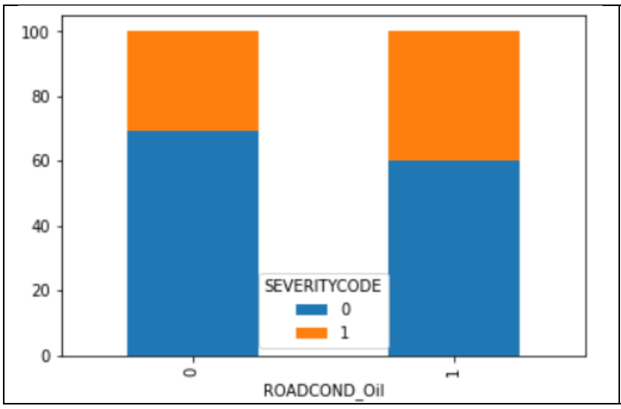


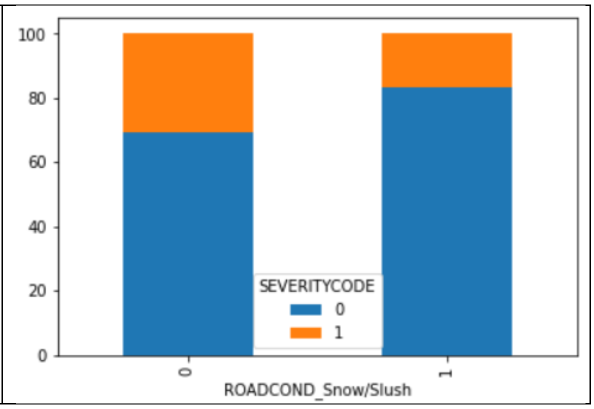
Figure 4.1: Imbalanced class distribution

There are 56,625 (30.96% of total) injury cases. Since the imbalanced class distribution may lead to unreliable machine learning result, oversampling technique is employed.

Relationship between several attributes and target variables are investigated.







5 MODELLING

Since government servants may not understand complicated model, both decision tree and logistic regression algorithms are chosen. They have great interpretability and can be easily understand by government servants.

5.1 Decision Tree

We use entropy to construct a decision tree with max depth of 20. Due to possibility of disruption of imbalanced data, area under ROC curve instead of accuracy is used to choose the optimum decision tree model.

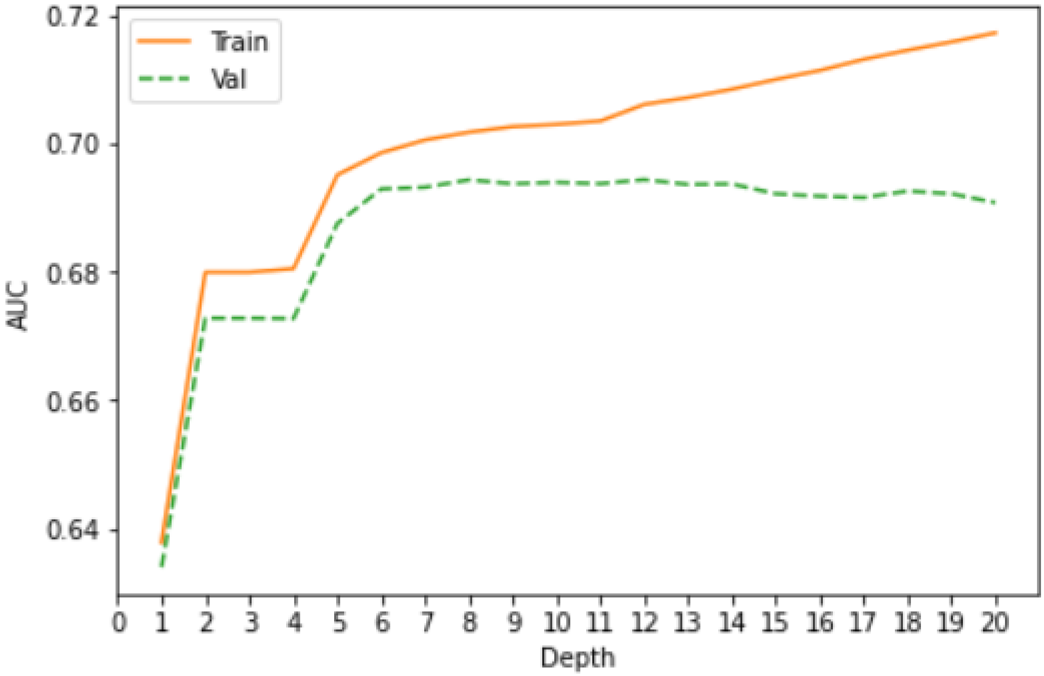
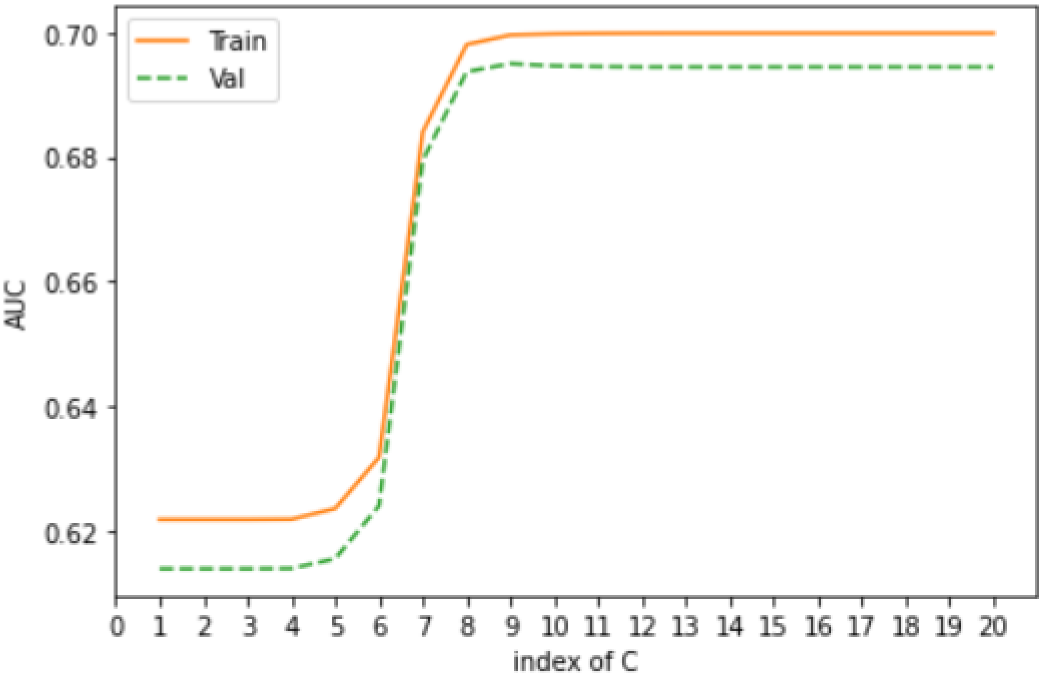


Figure 5.1: Decision tree by depth

Based on Figure 5.1, the best AUC generated by validation dataset is 0.694 with depth of 12.

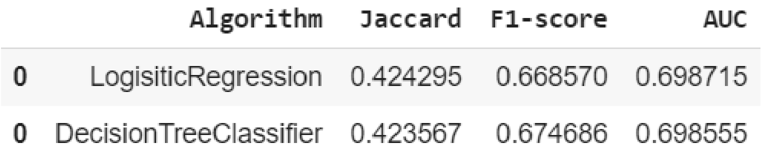
5.2 Logistic Regression

We construct logistic regression using different regularization strength. Similarly, area under ROC curve is used to determine the performance.



The best AUC generated by validation set is 0.6950 with C = 0.0264.

6 RESULT



	Algorithm	Jaccard	F1-score	AUC
0	LogisiticRegression	0.424295	0.668570	0.698715
0	DecisionTreeClassifier	0.423567	0.674686	0.698555

Both logistic regression and decision tree show similar result. Logistic regression is chosen due to its simplicity and interpretability. Next, we will delve into the importance of features. It may help the Seattle government to pinpoint the primary causes affecting the severity of an accident and take appropriate action to reduce the impact.

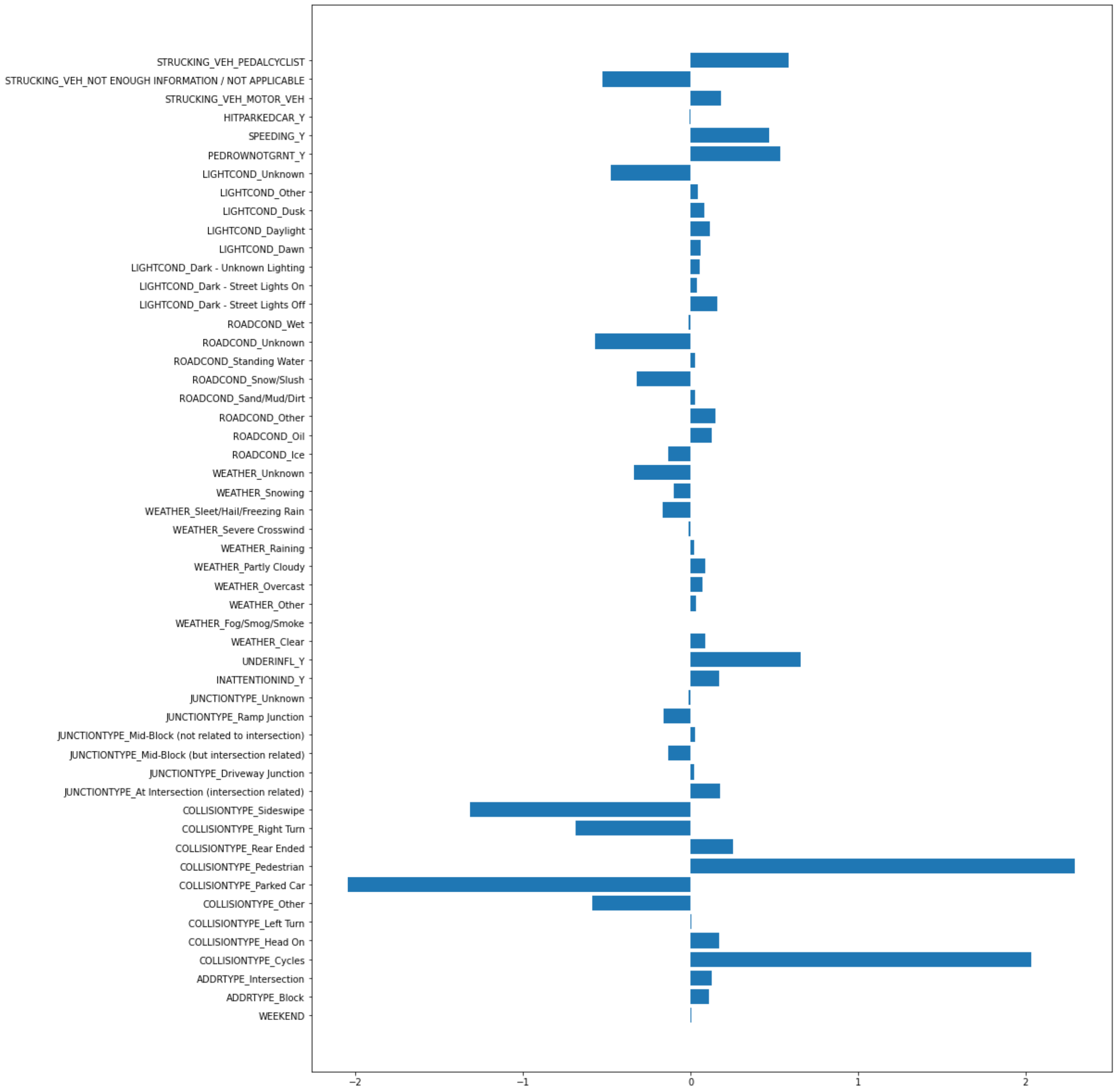


Figure 6.1: Variable Importance

7 DISCUSSION

Based Figure *6.1*, apart from collision type and strucking vehicles (out of Seattle government control), the primary risk drivers are as follows:

1) Driver under influence of drug or alcohol
2) Snowing/ Snowing road conditions
3) Speeding
4) Granted pedestrian right of way

During snowing weather, the visibility is poor. Limited visibility increases the chance of an accident. The problem could be solved by using de-icer to combat ice and snow from vehicle in a environmentally friendly way. With the slippery snowing road, a vehicle takes more time to completely stop the vehicle. Following too closely will result an accident. This situation is worsened if the vehicle has poor braking system. Seattle government may establish weather forecast platform to remind the driver to slow down and stay safe.

The remaining three risk drivers can be grouped as human factor. Government may amend the penalty for driving under influence of drugs and alcohols. For example, increase of minimum penalty and longer jail terms. Similar approach can be done for speeding.

8 Conclusion

In a nut shell, accident severity model has pointed out that majority of accidents are mainly due to human carelessness and irresponsibility. It is unforgiveable to those who build their happiness (speeding and consuming drugs and alcohol) at the expense of other’s life. Hence, Seattle government should introduce stiffer penalties to those driving under influence of drugs and alcohol, also speeding.