Car Accident Severity Analysis: Seattle, Washington (IBM Capstone Project)

The Car Accident Severity project aims to understand the effects of various factors on the likelihood and severity of car accidents using a Machine Learning Model.

Introduction

Road traffic injuries are among the ten leading causes of death worldwide, and they are the leading cause of death among young adults aged 15–29 years. These accidents also lead to 20–50 million non-fatal injuries, and many people incur a disability as a result of their injury. According to WHO, 1·25 million people worldwide died in road traffic accidents in 2013.
The world also suffers greatly on an economic front due to road accidents and the costs of these accidents are covered by taxpayer money. Among all countries, the USA has the largest economic burden of road injuries of $487 billion, followed by China ($364 billion), and India ($101 billion); according to a research journal published by THE LANCET.

Major Stakeholders

- Travelers
- Insurance Companies
- State Health Department
- Emergency Services
- Infrastructural Development Authorities
- Families of the Travelers
- Taxpayers

Problem

There is a lack of awareness amongst travelers regarding the risks they might be facing while taking certain routes, crossing certain areas, driving at a specific speed, driving on a specific road, and being inattentive while driving, etc. High-accident-prone areas are seldom inspected with regards to road maintenance, and deployment of additional emergency services personnel, causing additional damage.

Goal

This project aims to predict whether an accident that happens under a specific set of circumstances will be an accident limited to *property damage* or if it will include some form of *physical injury* to the driver and/or the passengers.

Data

The dataset that being used for this project is majorly provided by the government and pertains to the city of Seattle, Washington. It includes observations from 2004 to 2020. The number of observations in the data are enough to formulate a machine learning model. A large majority of the feature-set contains qualitative and categorical data, which is why performing a simple *Multiple Linear Regression* or* Polynomial Regression* is not the good option. The target variable for this model is the *level of severity* of the car accident (property damage only versus physical injury).

This is the dataset in CSV format in case it needs to be viewed: (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv)

After initial data exploration, we determined the following features to be most relevant when predicting *Accident Severity*.

Selected Independent/Predictor Variables

Feature Variables	Description
INATTENTIONIND	Whether or not the driver was inattentive (Y/N)
UNDERINFL	Whether or not the driver was under the influence (Y/N)
WEATHER	Weather condition during time of collision (Overcast/Rain/Clear)
ROADCOND	Road condition during the collision (Wet/Dry..)
LIGHTCOND	Light conditions during the collision (Lights On/Dark with light on)
SPEEDING	Whether the car was above the speed limit at the time of collision (Y/N)

Selected Target Variable

- **SEVERITYCODE**: A code that corresponds to the severity of the collision

Methodology

Data Collection

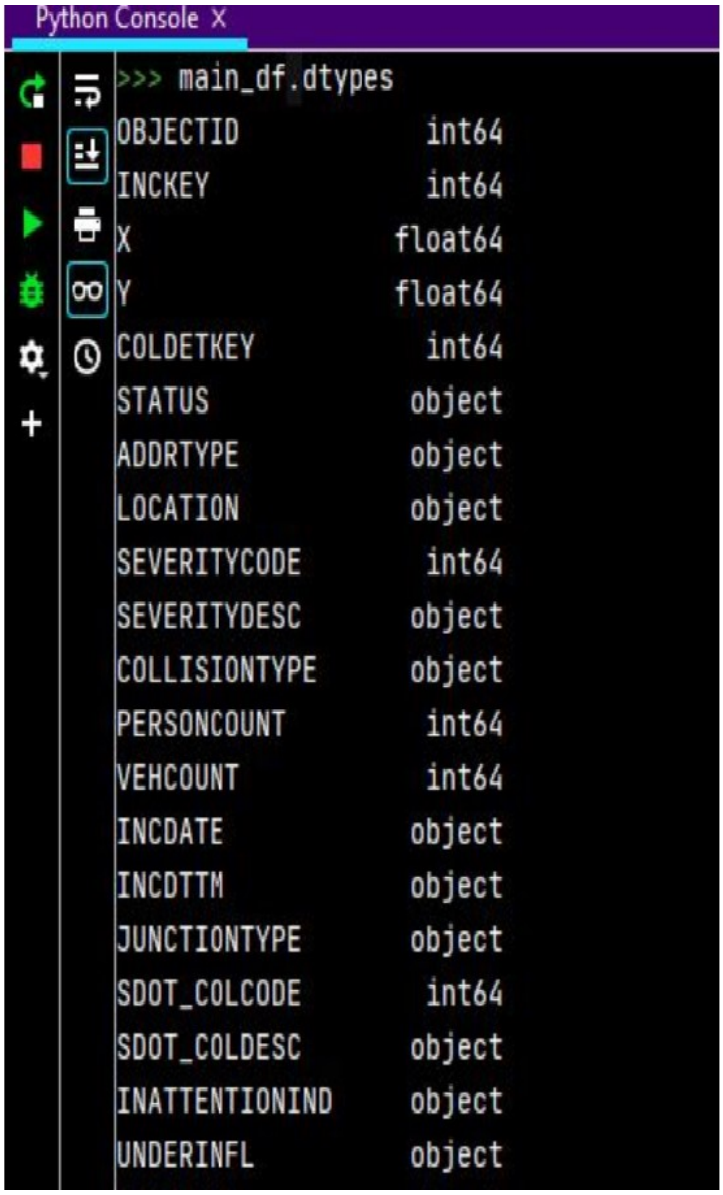
The dataset used for this project is a public dataset and illustrates the circumstances in which car accidents take place in Seattle, Washington, from 2004 to 2020.

Data Cleaning & Transformation

After gaining an understanding of the problem, the data had to be transformed to a form on which a machine learning model could be implemented. The first thing that was done was to check the data types of each variable and then explore how many variables were missing some entries.

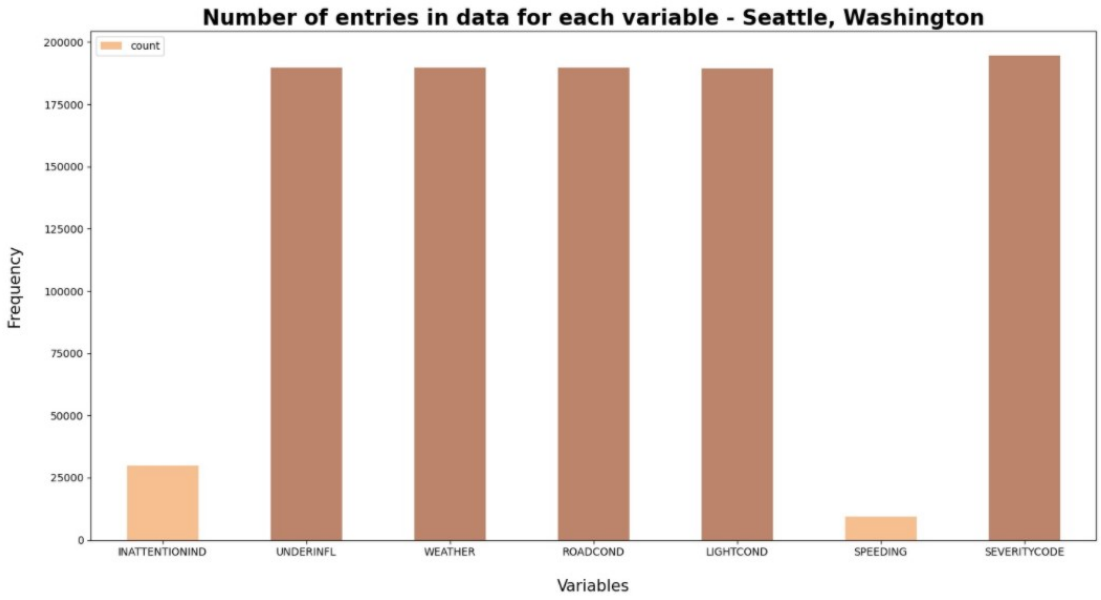
The data types in the Seattle dataset mostly comprised *categorical variables* and *objects*; it was concluded that a Simple/Multiple/Polynomial Regression would not work here. The variables in the dataset were listed in plain English and most of them were to be encoded with integers in an *ordinal* manner.

Data-types in the Dataset



The frequency of data points that contained entries that could be readily understood, for example, *Y* for *Yes*, *N* for *No*, or *0* for *False* and *1* for *True*, was higher for some variables than the others.

Frequencies in Dataset Columns before the Data was Transformed



After these analyses, it was concluded that some of the data will be dropped based on *materiality (will the dropped values significantly affect the analysis?)*, and the other will be encoded with integers. The *unknown* valuables were to be distributed back to the dataset in the same proportion the rest of the data was distributed, minus the *unknown* values.

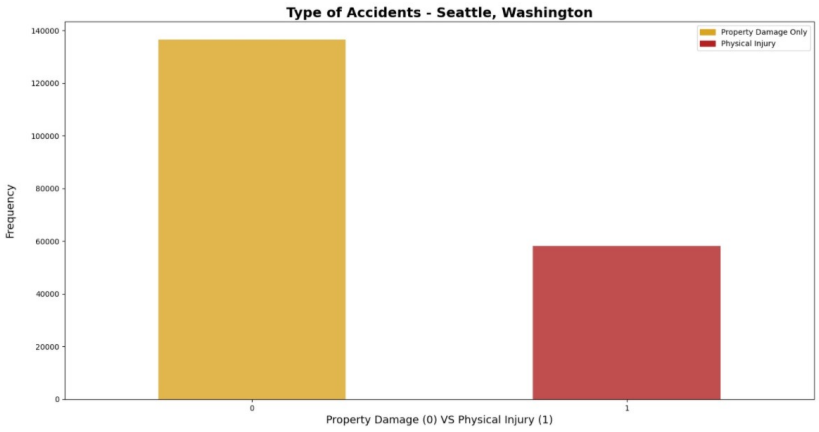
Subsequently, unavailable and unknown data points were re-distributed within the dataset in the same proportions as the known values in order to minimize the loss of data. Data that could not be salvaged was dropped, and the variables were encoded in integer forms; for example 0, 1, 2, and one unique identifier was retained for the dataset.

A new dataset by the name of "feature_df" was formulated after all the changes were made and relevant predictor variables were chosen through which a machine learning model would be created.

Next, the data was split into a *training set* and a *testing set* in order to train our model and test the predictions it makes in order to get accuracy metrics.

Most importantly, the number of accidents that were *property damage only* and the number of accidents including *physical injury* were compared in order to check the balance of the data so that biases could be minimized.

Balanced or Unbalanced?

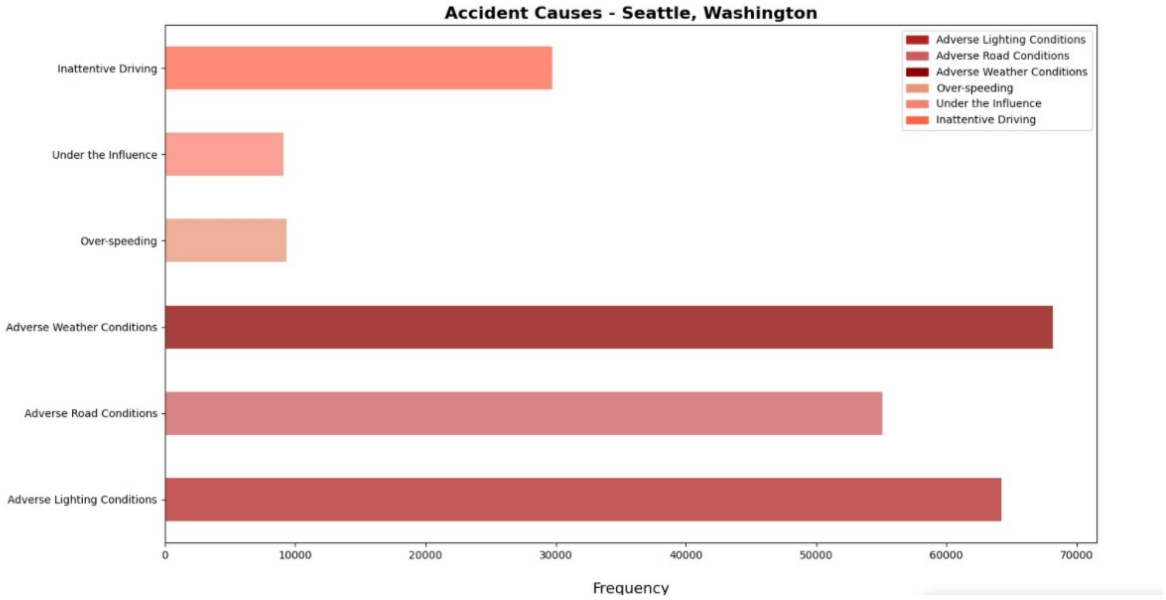


After recognizing that the dataset was clearly imbalanced, a Python library called *Imbalanced-Learn* was imported and *SMOTE* was used to *balance* the data to reduce the possibility of inaccurate predictions caused by having a significantly higher number of *Property Damage Only* data points within our *training set*. If this step was omitted, the model would have predicted a lot more *0s* or *Property Damage Onlys* than it should have.

Exploratory Data Analysis & Inferential Statistics

As a starting point, it was decided that any variable that is ~10% of the highest frequency variable that might cause an accident be included within the machine learning model and all 6 relevant variables fit this criterion. In order to check this, a bar graph was created and the frequencies were checked.

The Predictor Variables



Because the data was in a *cleaner* state now, it was easy to re-confirm that the *predictor variables* we had chosen by intuition were relevant whilst making the prediction and this was confirmed by the above visualization - most accidents had adverse conditions with respect to all the chosen variables.

Model Selection

Subsequent to gaining a complete understanding of the dataset, it was evident that there were no *continuous* variables and hence, a classification model was to be used instead of a *regression model*. There were four options that could have been implemented - *Decision Tree, K-Nearest Neighbor, Logistic Regression, and Support Vector Machine*.

Because SVM's training complexity is largely reliant on the size of the dataset and is not well suited to larger datasets, it was not used in this specific case.

In the end, *Decision Tree, Logistic Regression, and KNN* models were shortlisted as the machine learning classification algorithms that were to be tested.

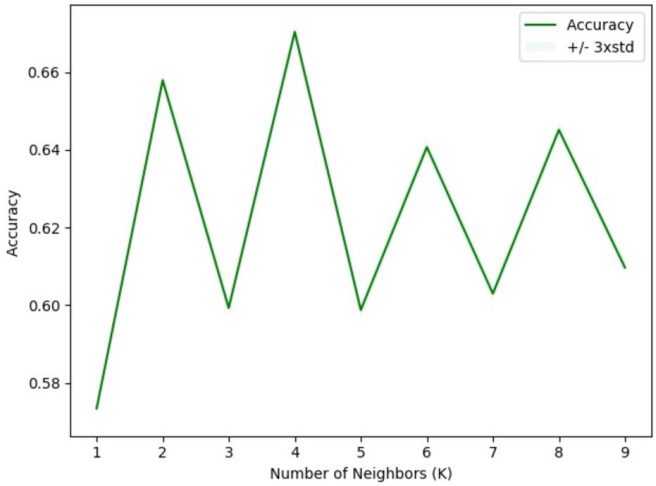
Results

The results of each of the three models varied; one excelled at predicting the *positives* accurately while the other predicted the *negatives* better. It was evident by the results that the predictions could have been improved if there was a more complete dataset at hand.

K-Nearest Neighbor

Before creating the KNN model, a loop from range 1 to 10 was run where the accuracy of the model was checked with varying values of K, and *K = 4* was chosen as it produced the highest accuracy. The result can be seen below:

***Choosing the right* K**



KNN Classification Metrics Report

	precision	recall	f1-score	support
0	0.93	0.70	0.80	52459
1	0.08	0.32	0.13	4343
accuracy			0.67	56802
macro avg	0.50	0.51	0.46	56802
weighted avg	0.86	0.67	0.75	56802

Decision Tree

Decision Tree Classification Metrics Report

	precision	recall	f1-score	support
0	0.73	0.71	0.72	40998
1	0.31	0.33	0.32	15804
accuracy			0.61	56802
macro avg	0.52	0.52	0.52	56802
weighted avg	0.62	0.61	0.61	56802

Logistic Regression

The Logistic Regression model tends to falter with larger datasets containing a high frequency of minority data points unless a more complex, penalty-oriented model is used. Surprisingly, it was also able to make fair predictions relative to the other two models.

Logistic Regression Classification Metrics Report

	precision	recall	f1-score	support
0	0.31	0.73	0.44	16878
1	0.74	0.32	0.44	39924
accuracy			0.44	56802
macro avg	0.52	0.52	0.44	56802
weighted avg	0.61	0.44	0.44	56802

It was noticed that the logistic regression model produced a higher-than-desired *uncertainty* which is illustrated by its *log loss*.

Log Loss

The value of *log-loss* is **0.69** for this model.

Discussion

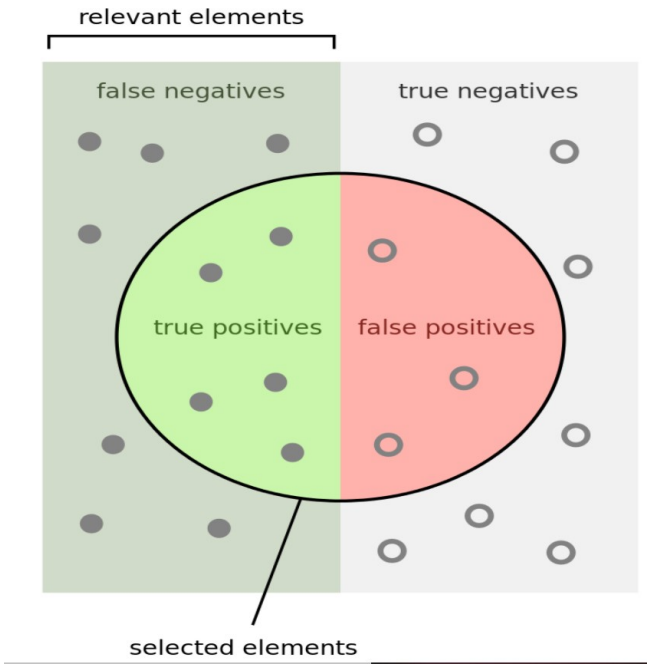
The results of all three machine learning models that were used varied significantly. One excelled at predicting the occurrences of *0* while the other would predict *0* and *1* with ~50-50 accuracies.

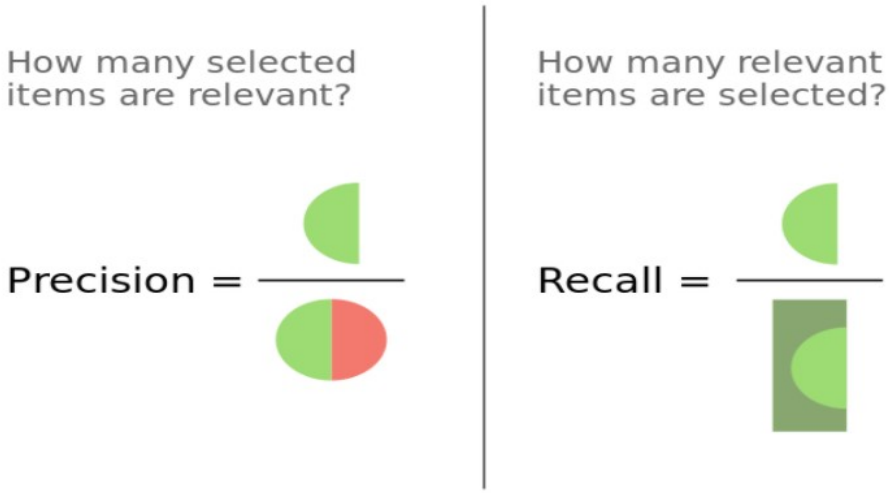
Comparison of Accuracy Metrics across the Models

Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.61	0	0.73	0.71
		1	0.31	0.33
Logistic Regression	0.44	0	0.31	0.73
		1	0.74	0.30
k-Nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

In order to understand what the report above means, it is necessary to take a look at what precision and recall signify.

Precision VS Recall





- **Precision:** How many selected elements are relevant?
- **Recall:** How many relevant elements are selected?
- **F1-Score:** Harmonic mean of *precision* and *recall*

Now that we have understood the different accuracy metrics used in the project, we can discuss what the varying results mean.

Comparing the Results

- **KNN:** This model has the highest *weighted average F1 Score*. It is highly accurate while predicting *negatives* but performs poorly when it has to predict the *positives*.

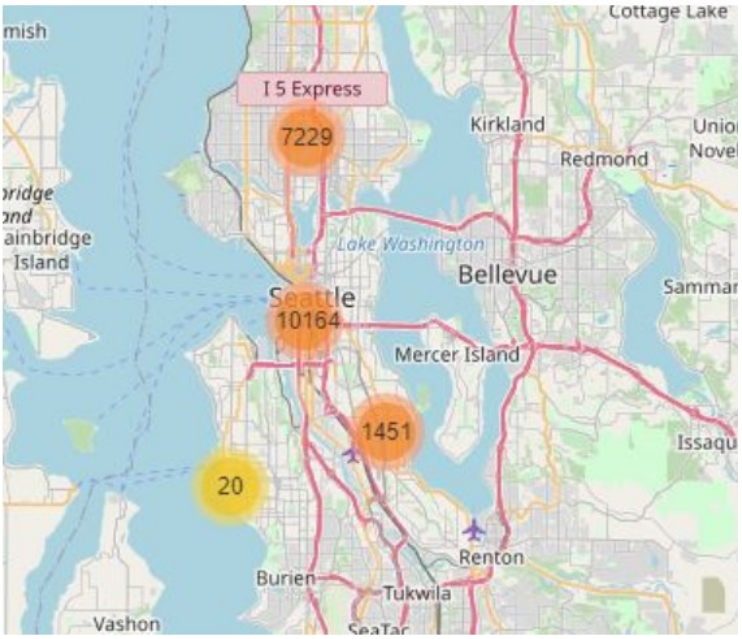
- **Decision Tree:** This algorithm predicts in a more balanced manner than KNN. Like *KNN*, it also predicts the *0s* with a higher degree of accuracy, but it can also predict the *1s* with a greater accuracy than the KNN implementation.
- **Logistic Regression:** This machine learning classifier has the lowest *weighted average F1 score* at *0.44* but it has the most balanced prediction. It classifies both *0s* and *1s* similarly and has the same *F1 score* for both.

Conclusion

It is evident by the models' accuracy ratings that there is still room for improvement. I believe that better models could have been made if the data at hand was more comprehensive and had less *unknown* and *missing values*. The analyses done could also have held more value if greater *target variable* class-data was available and was not limited to *property damage only* and *physical injury*. That way, whether an accident would happen in the first place could have been better predicted. It would also have possibly meant that the algroithms implemented would have given more accurate predictions as the correlations between *predictor* and *predicted* variable could have been stronger.

Recommendations

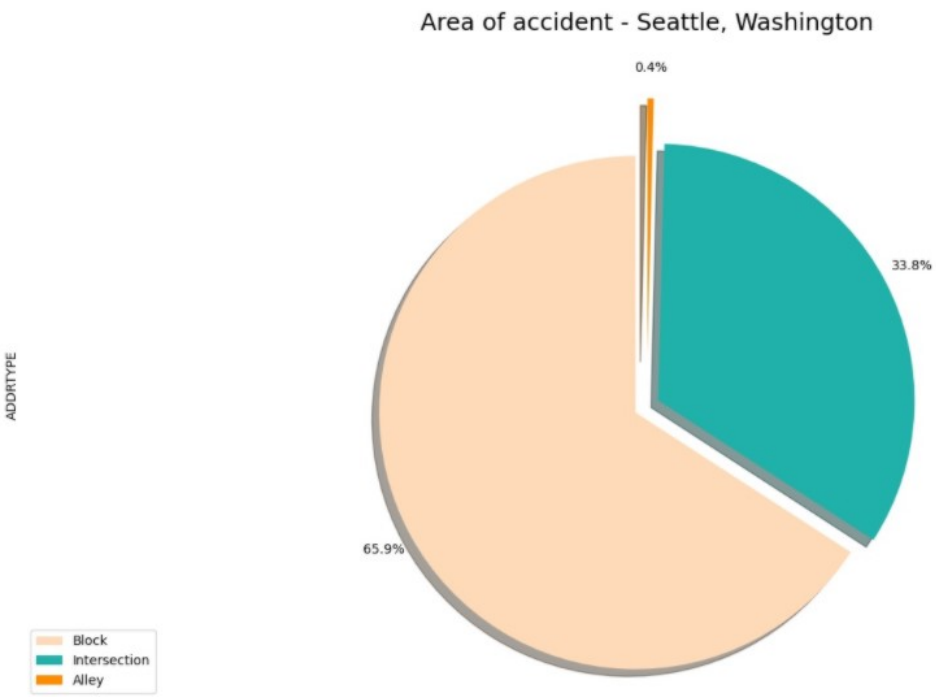
The I-5 highway and Central Seattle: The* Danger Zones*?



For Travelers & their Families

- Most accidents include adverse weather, road, and lighting conditions. The driver should make sure that they have a backup plan for tackling adverse weather conditions, e.g. a place to stay on their way if the weather forecast is not favorable. If possible, reschedule the day on which you travel. It is also advisable to travel during daytime in order to avoid accidents caused by hampered visibility.
- Inattentive driving, over-speeding, and being under-the-influence are also contributors to the likelihood of an accident occurring and closely rival the above-mentioned factors. Make sure that the driver is well-rested, follows the speed-limit, and is not under-the-influence while traveling.
- Drive carefully in central Seattle and the I-5 highway as that is where the most accidents happen.

Blocks & Intersections account for ~100% of Car Accidents in Seattle: is better traffic control needed?



For the Government

- Ensure that there are enough emergency responders in the areas where accidents happen the most. Also decrease the *lead times* of the responses in order to save more lives and keep law and order in check.
- Make sure there are hospitals near to accident hotspots in case of crashes including *severe physical injury*.
- There should be increased investment in *accident hotspot areas* and road & lighting conditions should be improved. Installation of a greater number of caution & safety signs should also be considered.
- Better traffic management systems can be implemented in order to avoid accidents at *intersections*.

For Insurance Companies

- Check to see whether the client that is being insured travels on the I-5 highway regularly. Also check how frequently they travel to/within Central Seattle.
- Check the weather, lighting and road conditions of where the client frequently travels to and where they live.