Car Accident Severity: Seattle, Washington

(Applied Data Science Capstone – IBM)

This project aims at understanding what factors play a vital role in the severity of car accidents in Seattle, Washington using Data Science Toolkit and predicting them prior to take necessary measures to avoid them using Machine Learning techniques.

Name: Vedant Mane

Project: Car Accident Severity Prediction

Course: Applied Data Science Capstone

Specialization: IBM Data Science Professional Certificate

Contents:

1. Introduction:

1.1. Background:

Seattle, also known as the Emerald city, is the largest city in both the state of Washington and the Pacific Northwest region of North America. It is home to a large tech industry with Microsoft and Amazon headquarters in its metropolitan area. The city has urban population of over 3.4 million as reported by PopulationStat. [1] As reported by curbed.com [2] in 2017, the total number of personal vehicles in Seattle in 2016 is approximately 435,000. The car population has more than doubled in the Seattle area since 2010. This tremendous increase in the number of vehicles has lead to higher number of accidents on the road explained merely by a simple probability. Worldwide, approximately 1.35 million people succumb to death due to road crashes every year, on average a total of 3,700 people lose their lives everyday in the road and an additional 20-50 million people suffer non-fatal injuries, often resulting in long-term disabilities.

1.2. Problem:

The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm can cost up to $871 billion in a single year. According to the WSDOT data, a car accident occurs every 4 minutes and a person dies every 20 hours due to a car crash in the state of Washington. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of car accidents can be reduced based on a few factors.

1.3. Stakeholders:

The reduction in the car accidents can be beneficial to several government bodies that work towards improving those road factors and car drivers themselves who may take precautionary measures to reduce the severity of the accidents.

2. Understanding Data:

2.1. Data Source:

The dataset requires pre-processing so that the machine learning model classifies the prediction to a categorical variable. The dataset has a total of 37 features and 194673 observations with variation in number of observations for every feature. The total dataset has high variation in the lengths of almost every column in the dataset due to the missing data values. These values could have been beneficial for our prediction algorithm had the data been present. The Metadata information for the dataset can be found at the link here.

2.2. Data Pre-processing:

The model aims to predict the severity of the car accident, considering that, the variable "SEVERITYCODE" of Severity Code that was in the form of 1 (Property Damage only) and 2 (Injury Collision) were encoded as 0 (Property Damage only) and 1 (Injury Collision). The attribute "INATTENTIONIND" for Driver's Attentiveness at the time of accident was considered 1 were value was 'Y' and 0 where data appeared missing. Similarly, the attribute "UNDERINFL" was encoded as 0 for the values 'N', '0', 'Nan' while 1 for 'Y', '1'. In the case of "SPEEDING", the 'Y' value was considered as 1, while missing values were considered to have the value 0. The Light conditions were at the time of accident were categorized into 4 groups based on the type of light and time of day as Daylight, Dark with Street Lights, Dark without Street Lights & Dark with unknown lighting conditions. Similarly, Weather Conditions were categorized as Clear, Cloudy or Low Visibility, Windy, Pouring for Raining or Snowing or Sleet/Hail/Freezing Rain and Unknown for missing or unknown values. The condition of the roads was classified as Dry, Wet, Blocked or Unknown conditions. The "JUNCTIONTYPE" attribute was classified into 2 types based on whether the accident took place at the intersection or elsewhere. Then, dummy variables were created for all the categorical columns (ROADCOND, LIGHTCOND, WEATHER).

Further we will normalize this data using the StandardScaler function and fit this data for our dataset and transform the same for model building.

2.3. Feature Selection:

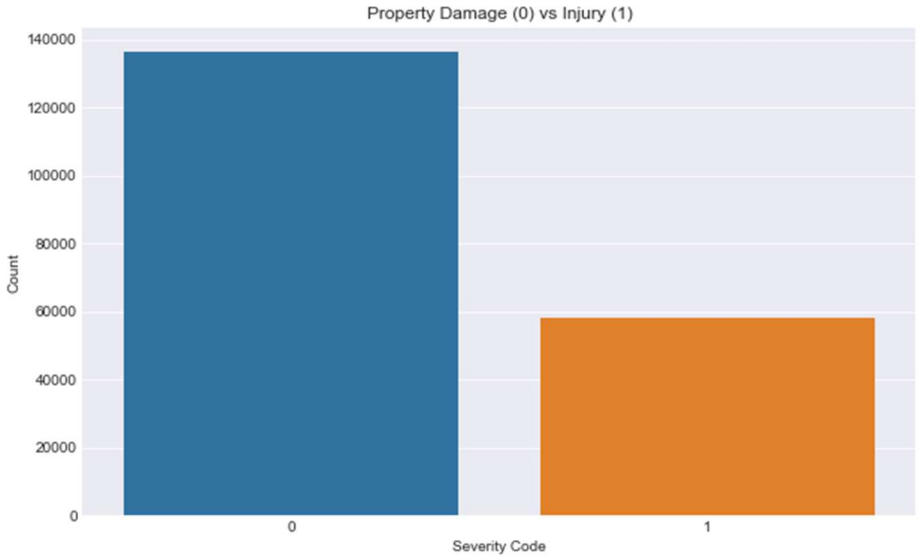
Feature	Description
Location (Address Type)	Description of the general location of the Collision.
Weather Condition	A description of the weather conditions during the time of the collision.
Car Speeding	Whether or not speeding was a factor in the collision.
Light Conditions	The light conditions during the collision.
Road Condition	The condition of the road during the collision.

Junction Type	Category of junction at which collision took place
Number of People involved	The total number of people involved in the Collision.
Number of Vehicles involved	The number of vehicles involved in the collision.
Driver's Attentiveness	Whether or not the Driver was attentive.
Driver under influence	Whether or not the Driver was under influence.
Severity Code	A code that corresponds to the severity of the collision: • 1 — Property Damage • 2 — Injury Collision

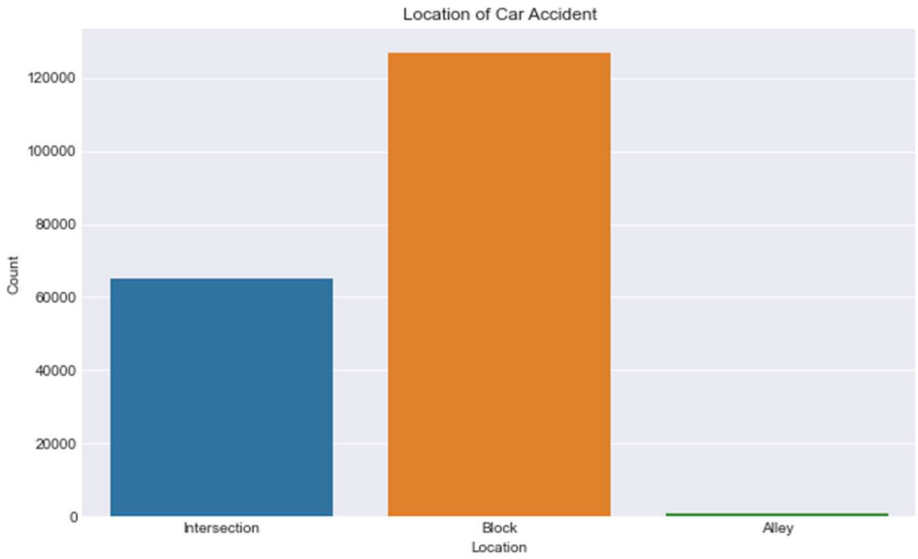
3. Methodology

3.1. Exploratory Data Analysis

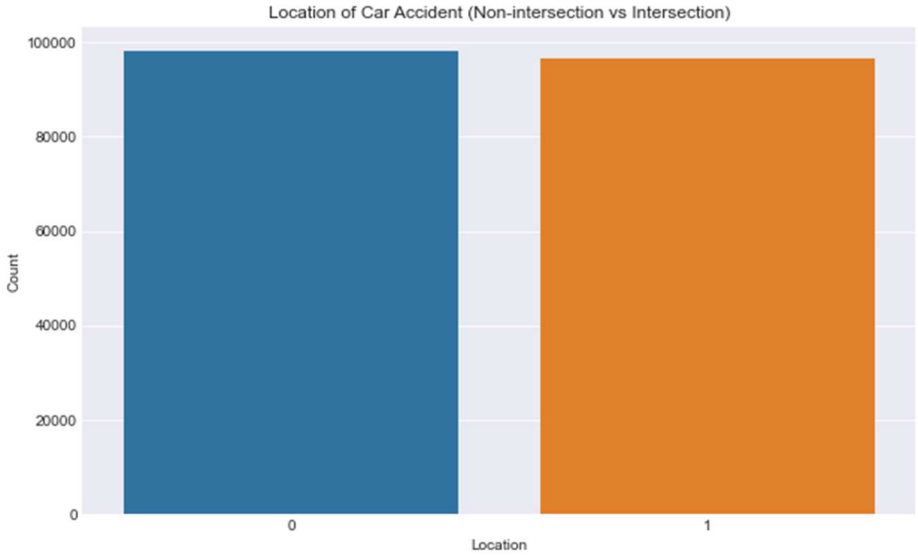
Every feature in our data i.e. transformed has categorical variables except for number of people and vehicles involved in the accident. From the dataset we can see that the data is highly skewed towards Property Damage than Injury. And hence, we will need to use the Imbalanced Learning package to remove bias from our Model.



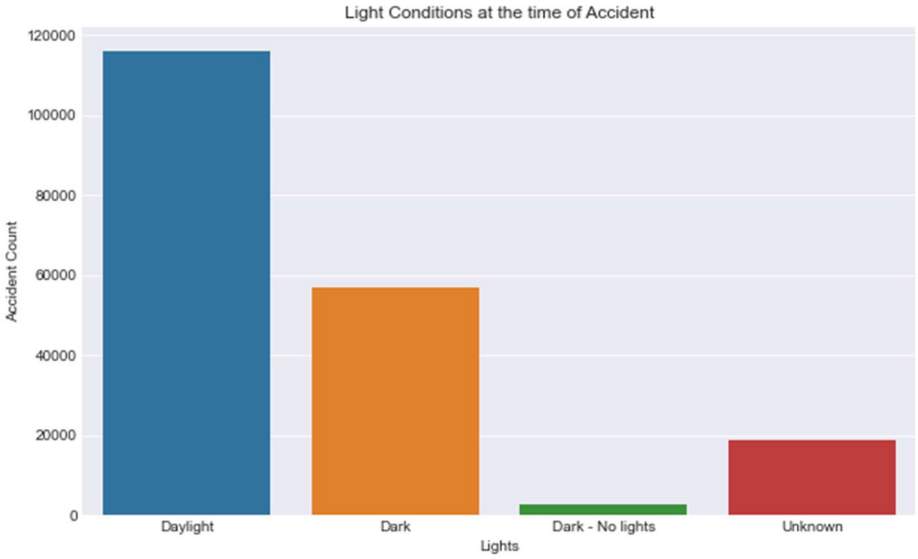
We can see from the below diagram that most of the accident take place at the Block, and almost all the accidents occur at either the Block or an Intersection.



We can infer from the below plot that the location on the road does not matter, the accidents occur almost same number of time at an intersection as they occur at other places.

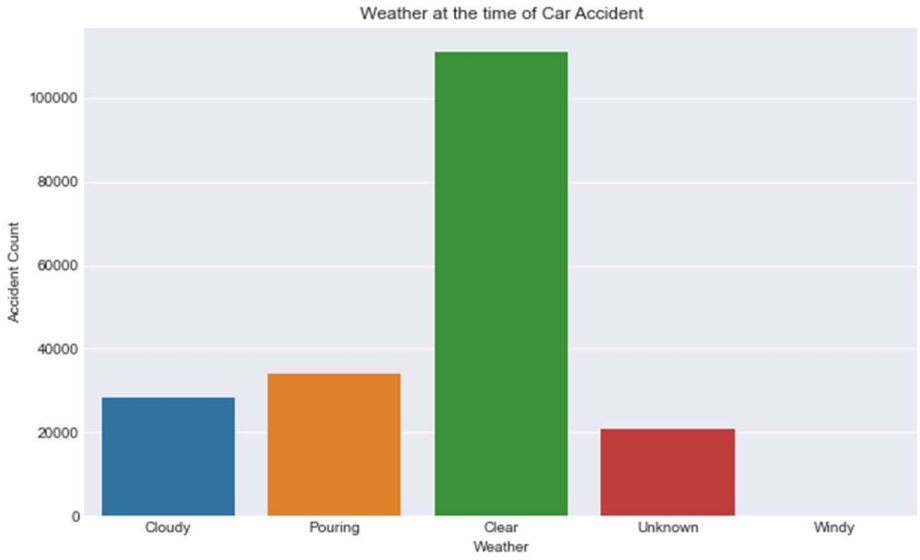


Most of the accidents take place during the daytime or when it is dark at places with proper street lights.

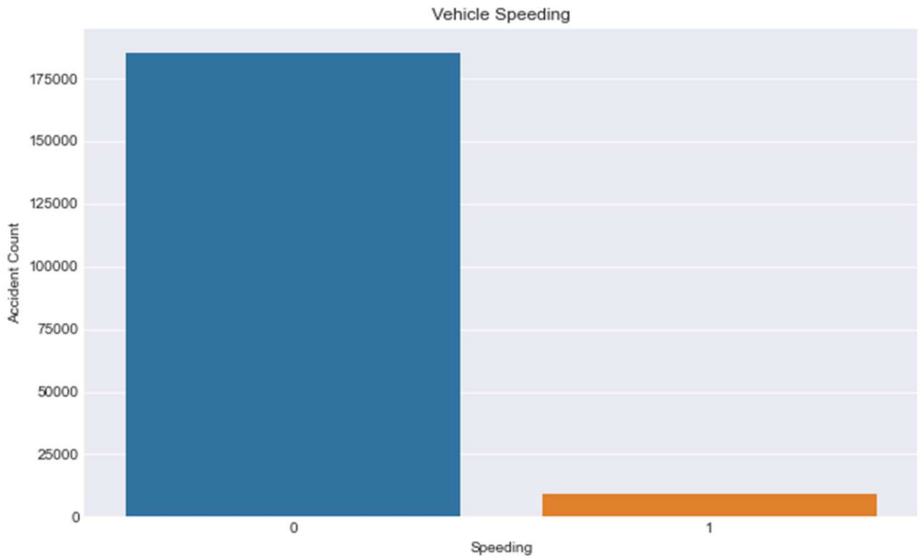


Improper Road Conditions hardly have a negative impact on the accidents as most of the accidents happen at places with dry roads. The next is followed by wet roads.

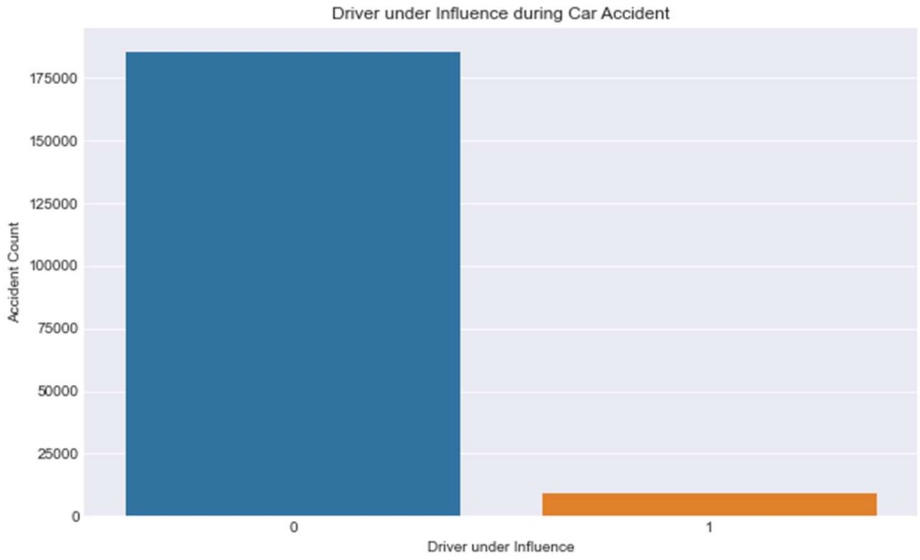
We can also see that most of the accidents occur when the weather is clear than at different conditions.



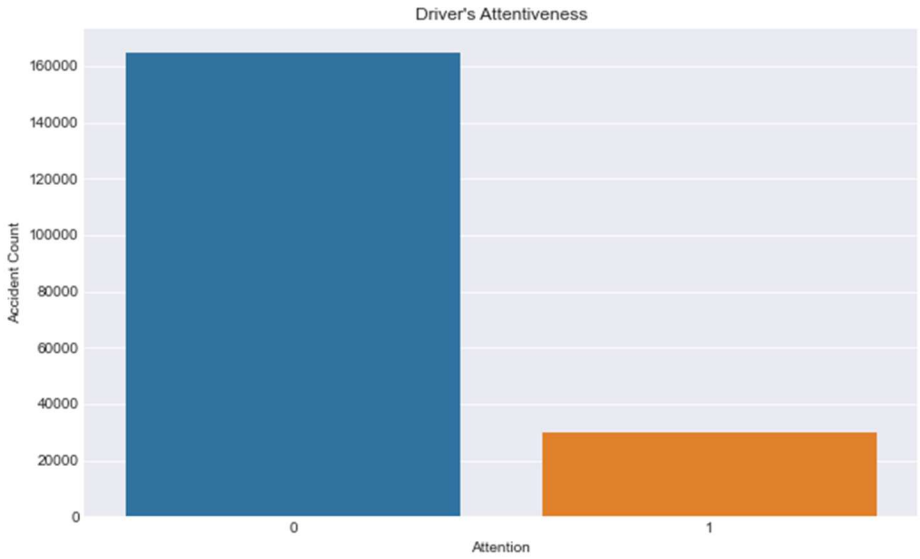
In this feature, we can see that Speeding does but cause much accidents but we have missing data in our column which is considered as not Speeding (0).



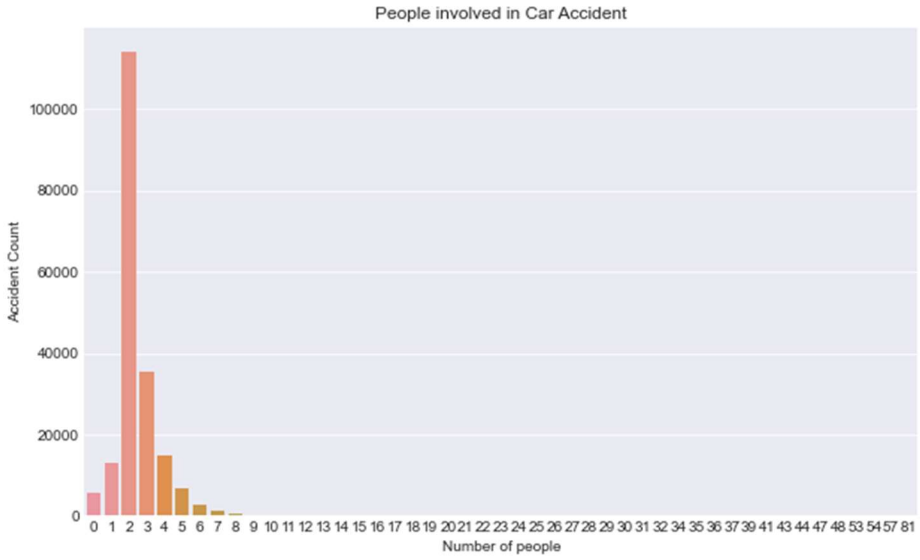
Similarly, we have very less data for Driver under Influence and cannot rely of this information. We have taken 0 for missing values.



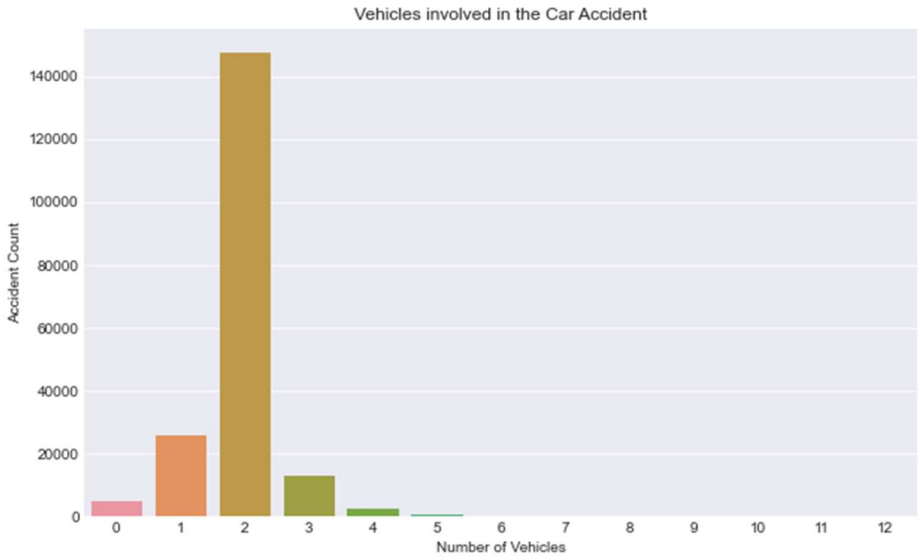
Same is the case for Driver's Attentiveness at the time of Car Accident.



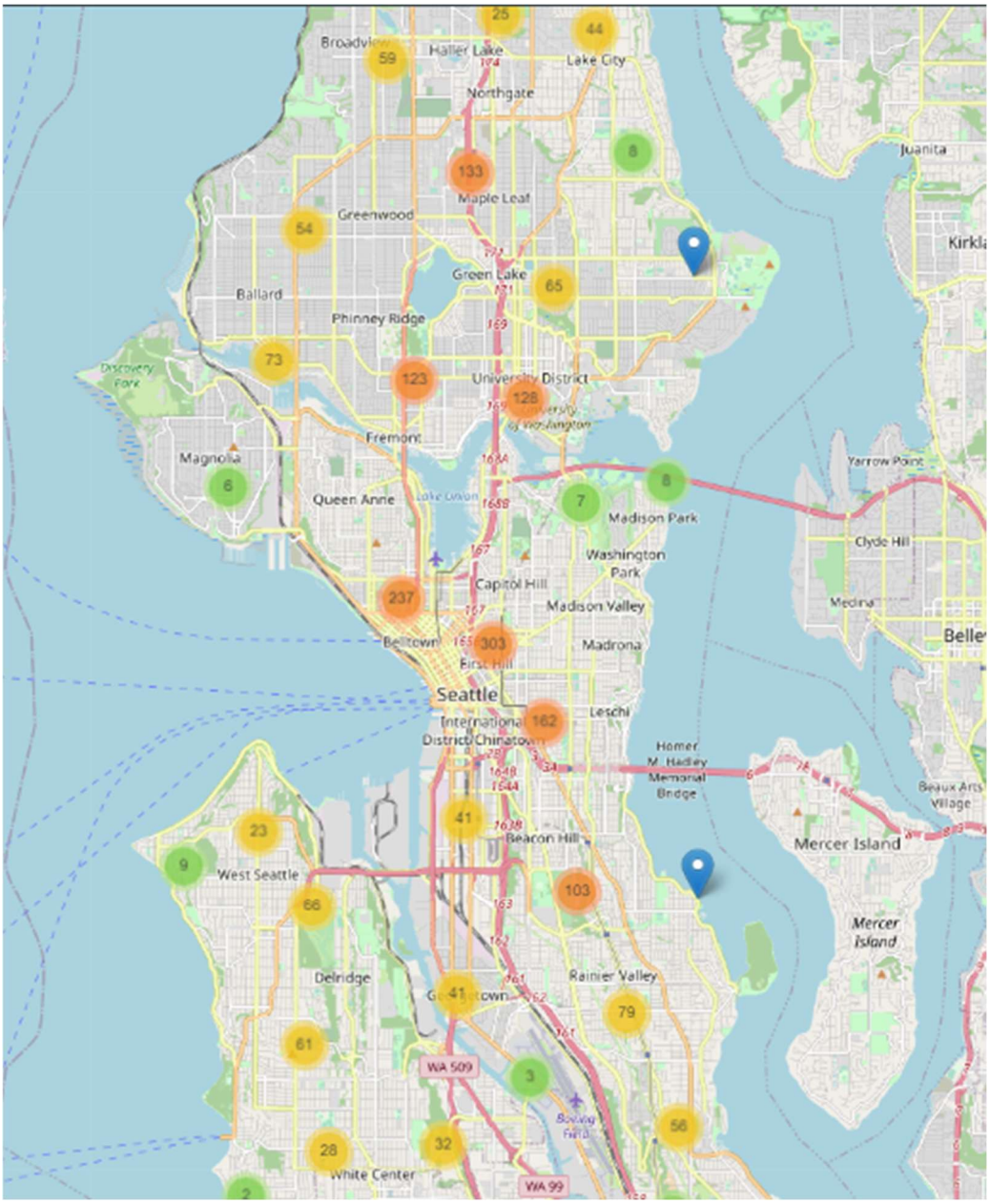
We can see that, in most of the cases, an average of 2 people are involved in an accident.



Also, for most of the accidents, 2 vehicles are involved.



We have plotted the first 10000 data values on the map of Seattle, Washington. We can see that most of the accident occurred at the main roads of the city, specifically near highways in the city's centre.



3.2. Model Selection

As this is a Classification Problem, we have used Logistic Regression and Decision Tree Analysis as the Machine Learning Algorithms for our dataset. Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependant variable. The Decision Tree Analysis breaks down a dataset into smaller subsets while at the same time an associated Decision Tree is incrementally developed. The final result is a Decision Tree with Decision Nodes and Leaf Nodes. Support Vector

Machines are inaccurate for large datasets. Also, KNN performs poorly when we have a high skewness in the dataset which is the case here.

4. Results

The dataset was split into training and testing sets and further balanced data was created using the SMOTE function from the imbalanced learning package.

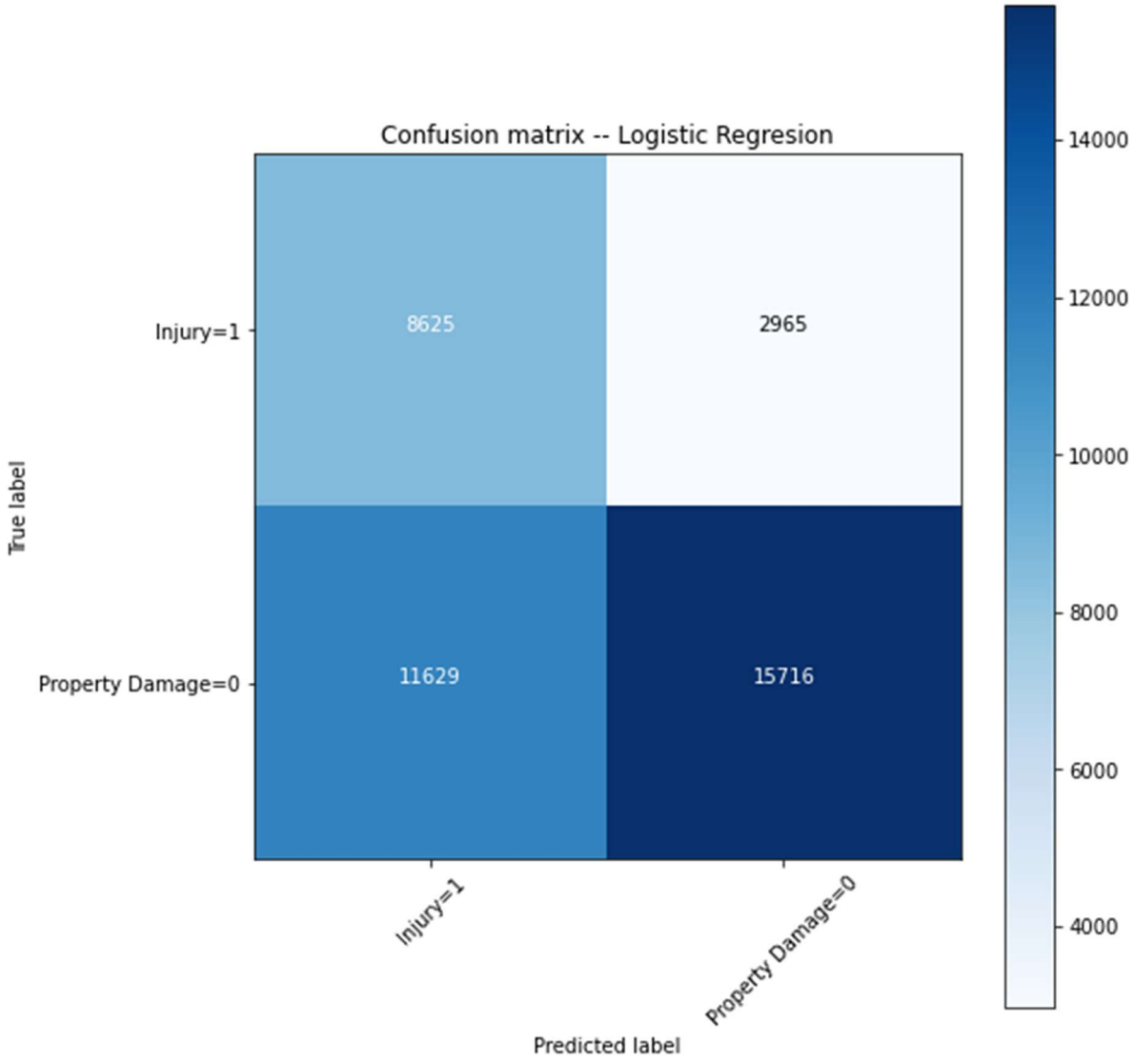
4.1. Logistic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification Model on the Car Accident Severity data. The C chosen for regularization strength was '0.01' with solver specified as 'liblinear'.

4.1.1. Classification Report

	Precision	Recall	F1 Score	Support
0	0.82	0.63	0.71	27345
1	0.43	0.67	0.52	11590
Accuracy			0.64	38935
Macro average	0.62	0.65	0.62	38935
Weighted average	0.70	0.64	0.65	38935
Log loss			0.63	

4.1.2. Confusion Matrix



4.1.3. Model Evaluation

Accuracy Score: `0.63850006420958`
F1 – Score: `0.5240912933220624`
Log Loss: `0.6311264004171047`

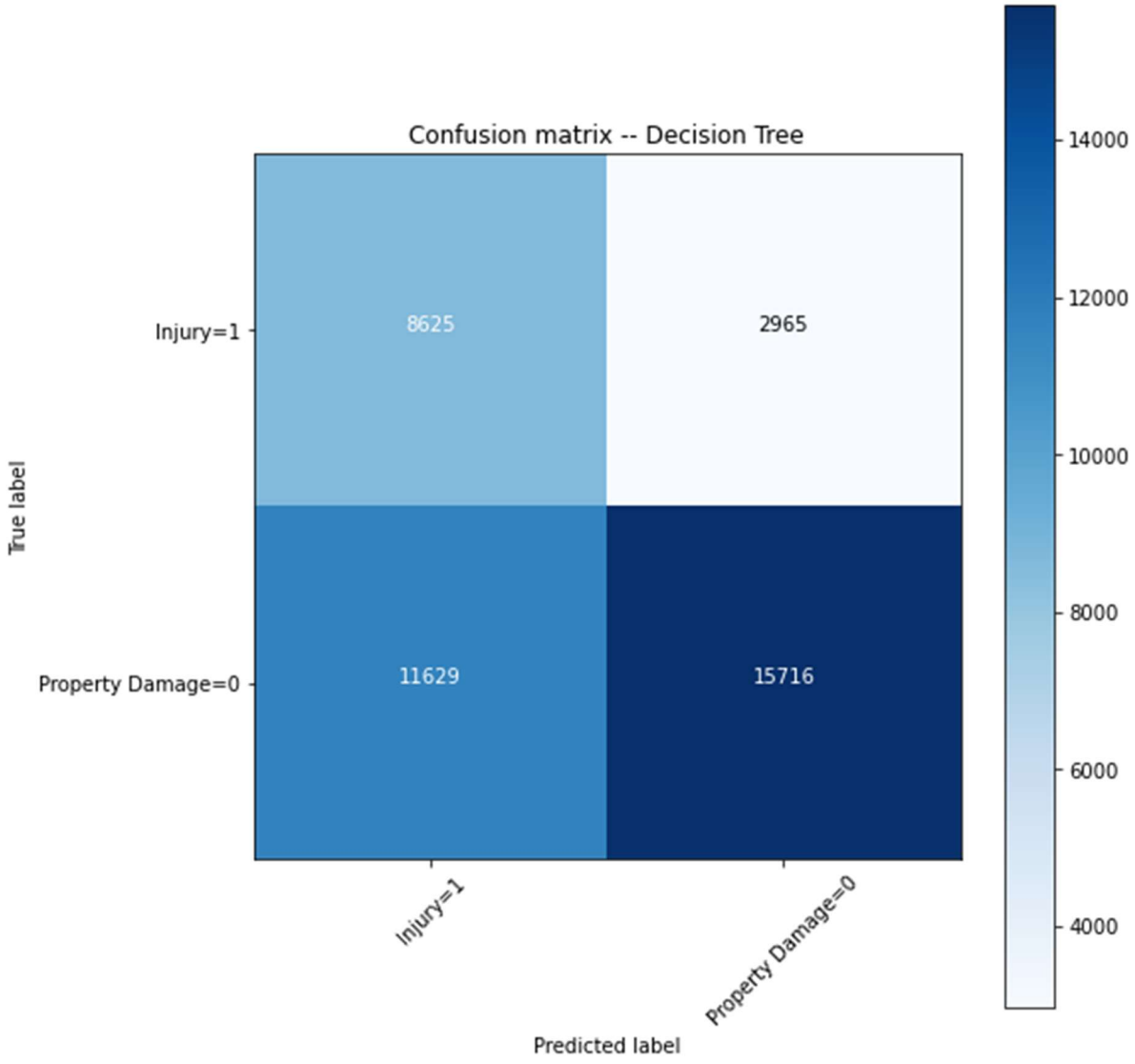
4.2. Decision Tree

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification Model on the Car Accident Severity data. The criterion chosen for the model was 'entropy' with a max depth of '4'.

4.2.1. Classification Report

	Precision	Recall	F1 Score	Support
0	0.84	0.57	0.68	27345
1	0.43	0.74	0.54	11590
Accuracy			0.63	38935
Macro average	0.63	0.66	0.61	38935
Weighted average	0.72	0.63	0.64	38935

4.2.2. Confusion Matrix



4.2.3. Model Evaluation

Accuracy Score: `0.6251701553871838`
F1 – Score: `0.5417033036050747`

5. Discussion

Algorithm	Average F1 Score	Property Damage (0) vs Injury (1)	Precision	Recall
Logistic Regression	0.65	0	0.82	0.63
		1	0.43	0.67
Decision Tree	0.64	0	0.84	0.57
		1	0.43	0.74

5.1. Average F1 Score

F1 – Score is a measure of the accuracy of the model, which is the harmonic mean of the model's precision and recall. The F1 – score shown above is the average of the individual scores of both the target variables. Here, when comparing the two models, we can see that their average F1 – scores are almost similar. However, average F1 – score does not depict the true picture of the model's accuracy because of the difference in precision and recall of both the models as our data is more biased towards the precision and recall of Property Damage (0) due to its weightage in the model.

5.2. Precision

Precision refer to the percentage of results which are relevant. The precision is calculated individually in order to understand how accurate the model is at predicting Property Damage and Injury individually. In terms of Precision, Decision Tree performs slightly better than Logistic Regression.

5.3. Recall

Recall refers to the percentage of relevant results correctly identified by the algorithm. Here, we can say that our Logistic Regression model performs better as it has balanced ratio of recall for both the target variables.

5.4. Recommendations

- Launch development projects for those areas with higher concentration of accidents and most severe accidents in order to minimize the effects of these two factors.
- Install safety signs on roads and ensure all necessary precautionary measures are taken by the people living in the area.
- Be extra careful around the highways along the city's centre as it has the highest number of accidents reported.

6. Conclusion

When comparing both the models, we can see that there is not much difference in their accuracy scores for the target variables. When looking at the two models individually, we can see that Decision Tree has higher precision however, Logistical Regression performs well when balancing the Recall for both the target variables. Also, the average F1 – scores are almost the same for both the models. Hence, it can be concluded that both the models can be used side by side for improving performance.

In retrospect, when comparing these scores with the benchmarks within the industry, it can be seen that they perform well but are not as good as the benchmarks. The models could have performed better if a few more things were available:

- A balanced dataset for the target variable
- More instances of the accidents recorded that took place in Seattle, Washington.
- Less missing values for features such as Speeding and Under Influence.
- More factors, such as precautionary measures taken while driving, etc.

7. References

[1] "populationstat.com," PopulationStat, October 2020. [Online]. Available: https://populationstat.com/united-states/seattle. [Accessed 03 October 2020].

[2] "seattle.curbed.com," Curbed.com, 10 August 2017. [Online]. Available: https://seattle.curbed.com/2017/8/10/16127958/seattle-population-growth-cars-transit. [Accessed 03 October 2020].