Car Accident Severity Analysis: Seattle

Submitted By: Srivastav Reddy Atla
Submitted on: 10/1/2020

Contents

1.Introduction:

Road accidents are one of the biggest problems in any nation, it leads to property damage or some cases loss of life as well. Many families in world are suffering because of road accidents.

In order to reduce the frequency of Car accidents in a community, an algorithm needs to be developed based on previous accident incidents that occurred in past.

Currently I'm trying to build an algorithm based on Seattle city accident data from police department showing all the accidents occurred from 2004 till present. When conditions are bad this algorithm model will alert the drivers to be careful.

This analysis will be helpful and have many real-world applications, this data analysis will be useful for the road transport authority to improve conditions of roads and to mitigate accidents in bad weather situations.

2.Data:

The data from Seattle PD consists of more than 190,000+ incidents which have been collected over past 15+ years. This data set has high variation in almost every column of dataset. The dataset has lot of empty columns, it could have been beneficial if the data has been present in it.

The model aim is to predict the severity of an accident, considering that the variable of severity code is in the form of 1 (Property Damage only) and 2 (Injury Collision) ,which I changed to 0 and 1 respectively for purpose of ease of representation. Changed the string values in other parameters to int, for light condition, Light was given 0 along with medium as 1 and Dark as 2. For Road Conditions, Dry as 0, Mushy as 1 and wet as 2. As for weather condition Clear -0, Overcast-1, windy-2, Rain/Snow – 3 was assigned to each variable.

Feature Variables	Description
INATTENTIONIND	Whether the driver was inattentive? (Y/N)
UNDERINFL	Whether the driver was under alchohal infulence? (Y/N)
WEATHER	Weather Condition during collison (clear, rainy,..etc)
ROADCOND	Weather Condition during collison (wet, dry,..etc)
LIGHTCOND	Light Conditions during collisions (Lights on, Dark with lights on.)
SPEEDING	Whether the car is over the speed limit or not

Table 1: Variable Description

The rows where the data is not proper and not complete , has been rremoved from the data frame. The is cleaning of data lead to almost loss of 5000 rows which had redudent data.

The Variables in above table is going to be our data for the algorithm model.

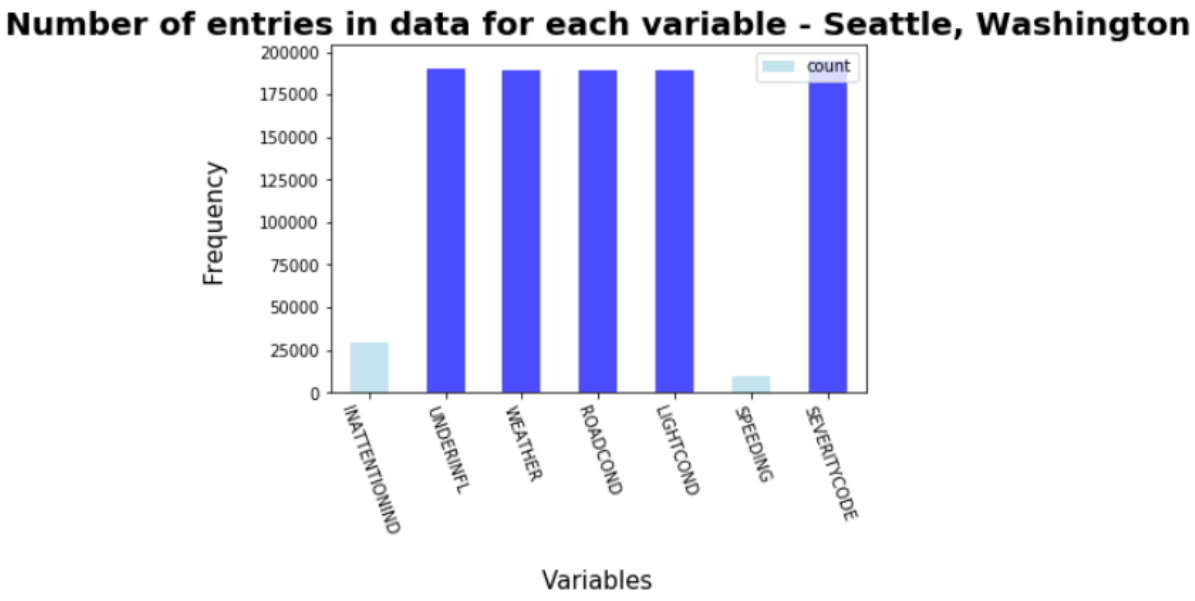

Fig1: Data Entries

3. Methodology

Exploratory Analysis

Considering the feature set and the target Variable are categorical Variables with the types of weather, road, light conditions being on above level 2 categorical variable whose values are limited and usually based on finite group whose correlation might depict a different image than the reality.

The above figure illustrates, after data cleaning has taken place, the distribution of the target variables between Physical Injury and Property Damage Only. As we know that the dataset is supervised but an unbalanced dataset where the distribution of the target variable is in almost 1:2 ratio in favor of property damage. It is very important to have a balanced dataset when using machine learning algorithms.

From the figure 1, it is clear and understandable that the Speed and Driving Under Influence are the two factors which contributed very less. The factor which had large number of accidents under adverse conditions was adverse weather conditions while adverse lighting condition had the second most number of accidents caused by it.

Machine Learning Models:

Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable

Decision Tree Analysis: The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

4. Results

The results are different for two models that we used, one worked predicting positives better while other predicted negatives better.

Decision Tree

The criteria chosen for the classifier was entropy and the max depth was 6.

Decision Tree Report:

```
Accuracy score for Decision Tree = 0.5760076740692476
```

	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

Table 2: Decision Tree Report

Decision Tree Confusion Matrix:

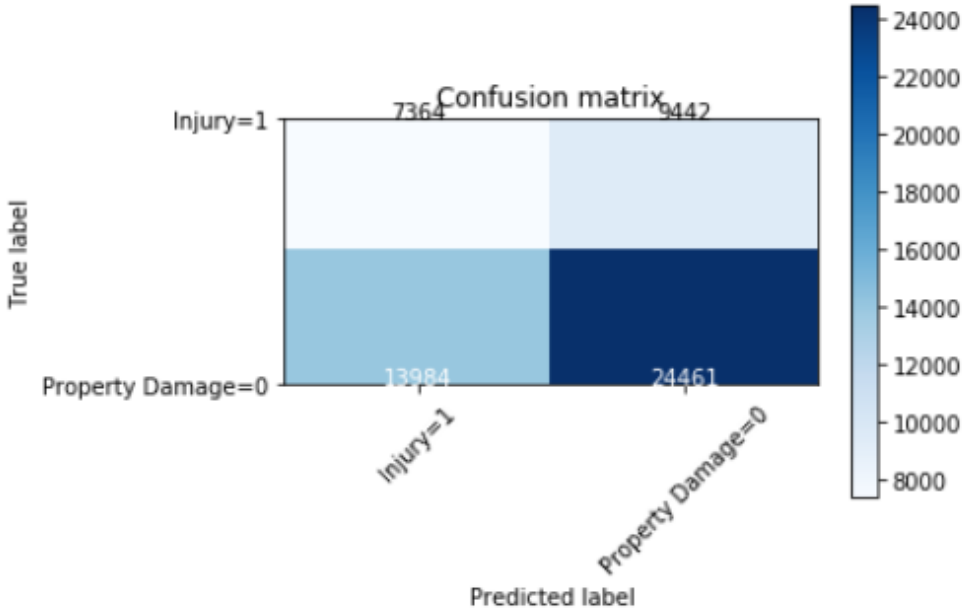

Fig 2: Decision Tree Confusion Matrix

Logistic Regression

Logistic Regression Classification Report

```
Accuracy of Logistic Regression 0.5888219217751715
```

	Precision	Recall	F1-score
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.61	0.59	0.60
Weighted Avg	0.68		

Table 3: Logistic Regression Report

Logistic regression confusion matrix

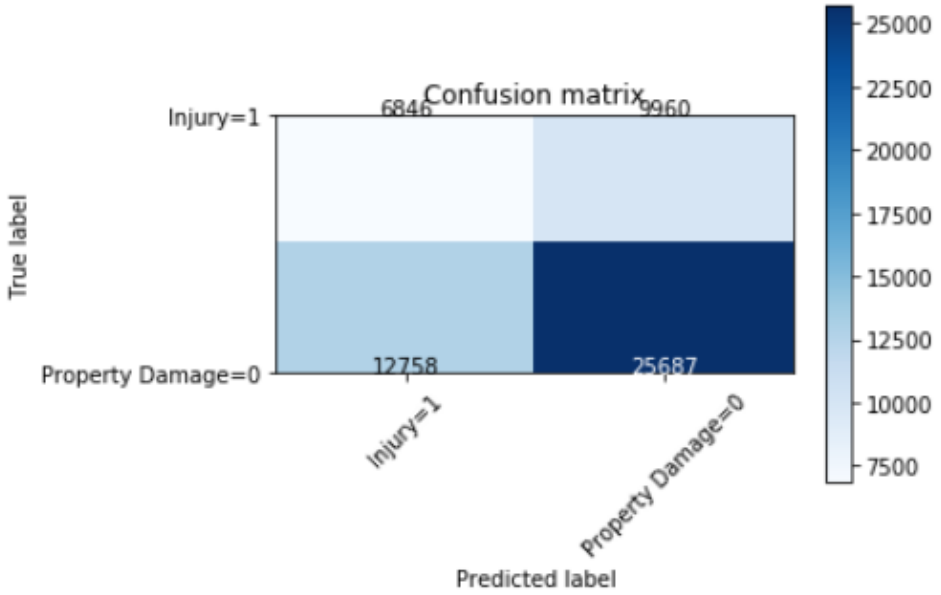

Fig 3: Logistic Regression Confusion Matrix

5. Model Accuracy

Precision: Precision refers to the percentage of results which are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive

Recall: Recall refers to the percentage of total relevant results correctly classified by the algorithm. It is calculated by dividing true positives by true positive and false negative

F1-Score: It is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0

Algorithm	Avg f1- score	Property Damage (0) vs Injury(1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41

6. Conclusion

When comparing all the models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the models individually as a whole and how well they perform for each output of the target variable. When looking at the two models, we can see that the Decision Tree has a more balanced precision for 0 and 1. Whereas, the Logistic Regression is more balanced when it comes to recall of 0 and 1. Furthermore, the average f1-score of the two models are very close but for the Logistic Regression it is higher by 0.04. We can be concluded that the both the models give equal performance.

7. Recommendations

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident.

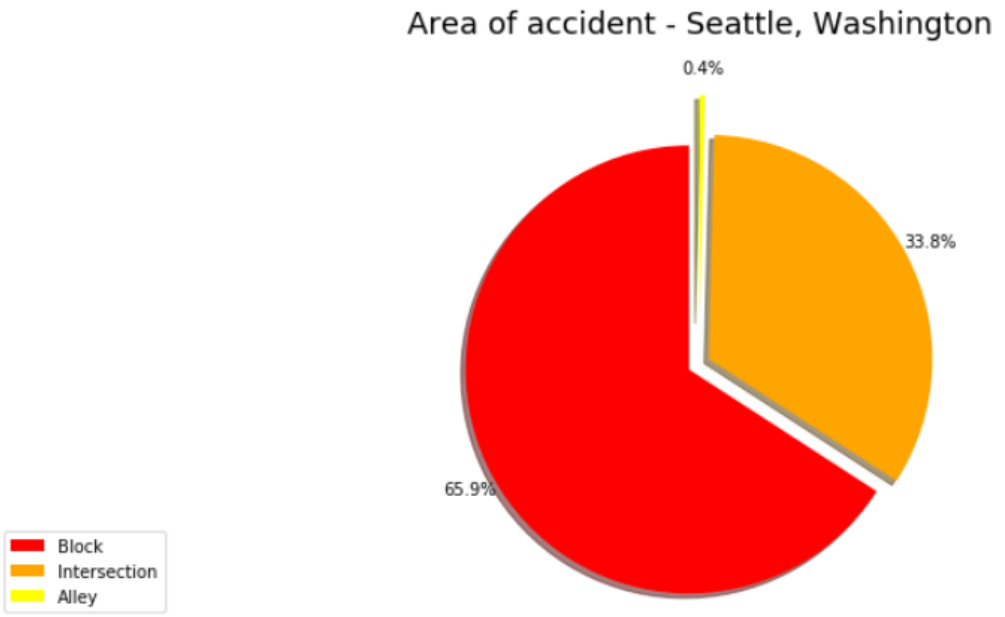

Fig 4: Accident Areas

Based on above chart it is evident that all the accidents happened at intersections and blocks. Seattle Public department can take measurements like installing more safety signs around those regions, increase traffic police patrolling and make sure that the people follow all safety signs.