Car Accident Severity Analysis

Kabileshwaran M

1. Introduction

Car accidents are one of the major problems that found across the world. The cause of car accidents may be due to lack of focus or due to natural reasons or even due to the other drivers. But most of the times it is due to negligence. The natural reasons will be mostly due to severe weather conditions such as heavy rain or heavy snowfall. The severity of car accidents are very high because there is a high probability of causing people's death.

The project aims to analyse severity on car accidents by reducing the chance of it's occurance. The project uses various machine learning techniques to analyse the data and find a best solution. It project's main stakeholders are the car drivers. The main question that this project will be addressing is,

"What are the main factors of the Car accidents and how the severity can be reduced?"

2. Data

2.1 Feature Selection

There are plenty of datasets based on car accidents are available from various sources. The one we used in the project is on the car accidents that taken place in Seattle city from 2004 to 2020. The dataset has it's first attribute of Severity Code. The Severity Code 1 means "Property Damage Only" and 2 means "Physical Injury". We will be encoding 1 and 2 with 0 and 1 for better understanding and simplicity.

The dataset contains many unknown data in many rows. So, we used the frequency of the data in occur in other rows and filled in the place of the unknown data. This involved in some loss in information gain from the start itself. The redundant data are cleaned and the new modified dataset will be taken for further analysis.

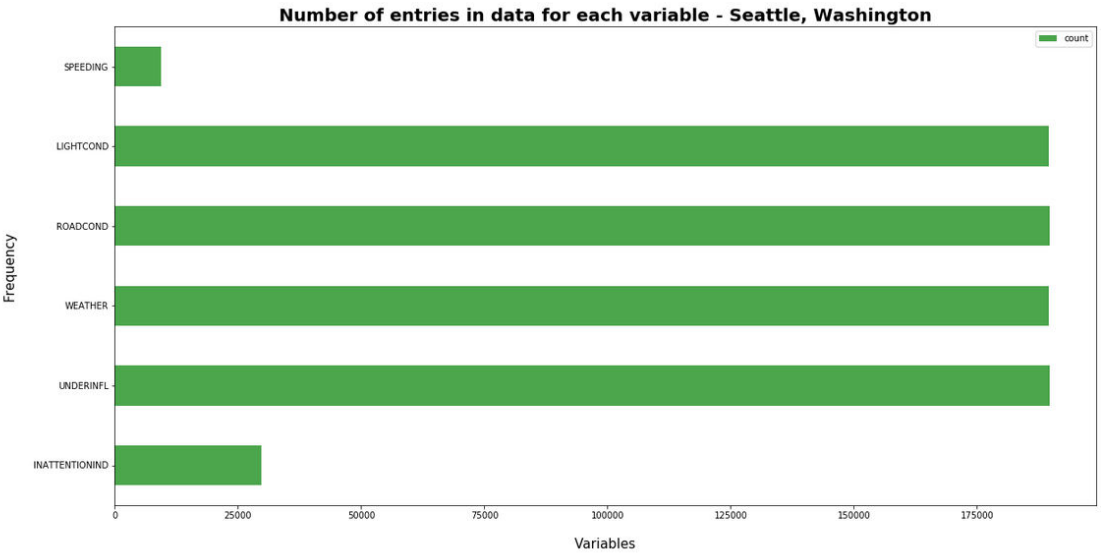
The feature variables that we took from the dataset for the project are,

- SPEEDING - Whether the car was above the speed limit at the time of collision
- ROADCOND - Road condition during the collision
- INATTENTIONIND - Whether or not the driver was inattentive
- UNDERINFL - Whether or not the driver is under influence
- LIGHTCOND - Light conditions during the collision
- WEATHER – Weather condition during the time of collision

3. Methodology

3.1 Exploratory Analysis

The feature set and target variable are categorical variables from the initial observation. The attributes we used here are weather, road condition, attentiveness of the driver, light condition, influence of the driver and speeding. These variables makes more impact in a car accident. Hence, these are selected for the exploratory analysis. The below chart depicts the frequency of the variables that causes the car accident.



It can be seen that the accidents due to weather, bad road conditions, bad light conditions and the driver under influence are very high in comparing with the accidents due to the inattentiveness of the driver during driving and speeding of the car.

3.2 Machine learning model selection

The machine learning models that are used for the project are,

- **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 final result is a tree with decision nodes and leaf nodes.

Note the Support vector machine model is not used in the project because it is less accurate with the large datasets. And it’s more suitable for texts and images.

4. Results

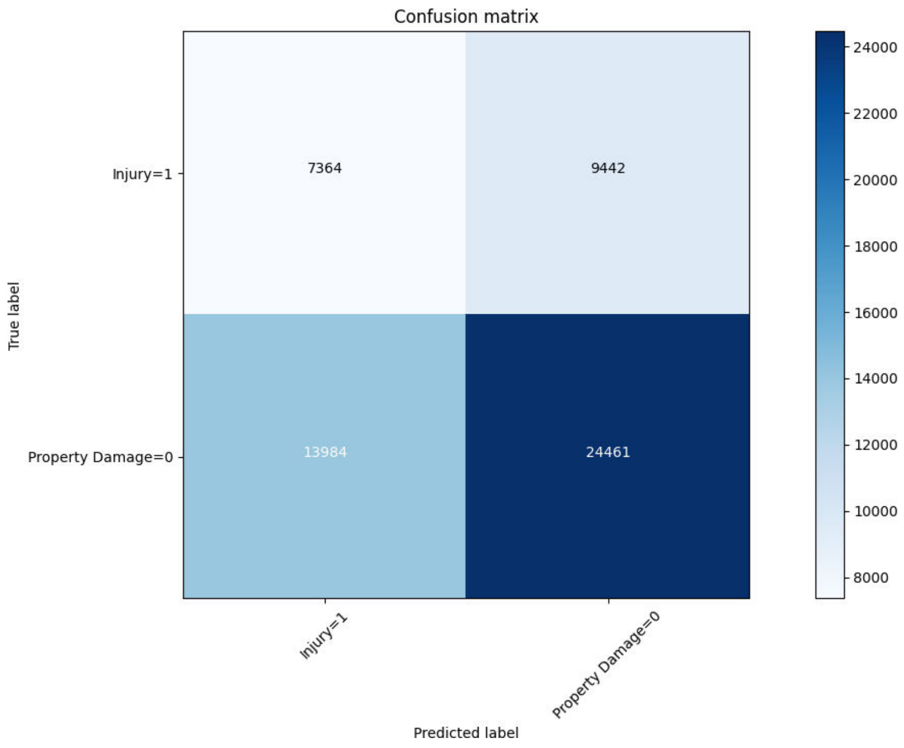
4.1 Decision Tree Analysis

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 classifier was 'entropy' and the max depth was '6'. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

4.1.1 Classification Report

	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

4.1.2 Confusion Matrix



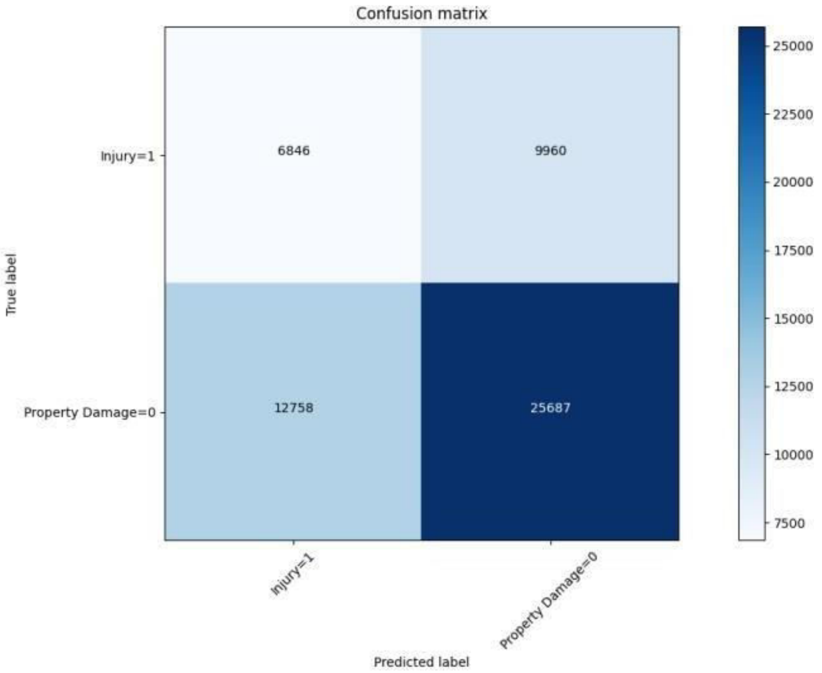
4.2 Logistic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data. The C used for regularization strength was ‘0.01’ whereas the solver used was ‘liblinear’. The post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier

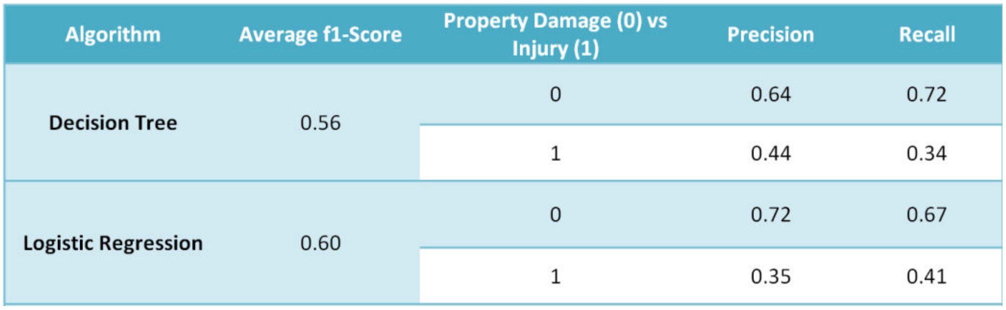
4.2.1 Classification Report

	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		

4.2.2 Confusion Matrix



5. Discussion



Algorithm	Average 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

5.1 Average f1-Score

f1-score 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.

The f1-score shown above is the average of the individual f1-scores of the two elements of the target variable i.e. Property Damage and Injury. When comparing the f1-scores of the two models, we can see that Logistic Regression has the highest f1-score meaning that it has a higher precision and recall. Whereas, the Decision Tree model's f1-score is the lowest at 0.56. The f1-score of the Logistic Regression is at 0.60 which can be considered as an above average score. However, the average f1-score doesn't depict the true picture of the model's accuracy because of the different precision and recall of the model for both the elements of the target variable. Hence, it is biased more towards the precision and recall of Property Damage due to its weightage in the model.

5.2 Precision

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive.

The highest precision for Property Damage is for Logistic Regression, whereas for Injury it is the Decision Tree. The Precision is calculated individually above in order to understand how accurate the model is at predicting Property Damage and Injury individually. For the Decision Tree the precision of 0 is 0.64 and for 1 it is 0.44 which is fairly good. As for the Logistic Regression model, for 0 it is at 0.72 and for 1 it is 0.35. In terms of precision, the best performing model is the decision tree.

5.3 Recall

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative.

The highest precision for 0 is when using the Decision tree model at 0.72 as for 1 it is the Logistic Regression model at 0.41. As for the Logistic Regression, the recall for Property Damage is 0.67 and for Injury it is 0.41. The recall for Property Damage and Injury is the most balanced in terms of being good for both the outputs of the target variable.

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 two models individually as a whole and how well they perform for each output of the target variable. When comparing these scores, we can see that the f1-score is highest for Logistic Regression at 0.60. 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. It can be concluded that the both the models can be used side by side for the best performance.

In retrospect, when comparing these scores to the benchmarks within the industry, it can be seen that they perform well but not as good as the benchmarks. These models could have performed better if a few more things were present and possible.

- A balanced dataset for the target variable
- More instances recorded of all the accidents taken place in Seattle, Washington
- Less missing values within the dataset for variables such as Speeding and Under the influence
- More factors, such as precautionary measures taken when driving, etc.

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, if any.

8. References

- *https://www.macrotrends.net/cities/23140/seattle/population#:~:text=The%20current%20metro%20area%20population,a%201.2%25%20increase%20from%202017*.
- *https://www.seattletimes.com/seattle-news/data/housing-cars-or-housing-people-debate-rages-as-number-of-cars-in-seattle-hits-new-high/#:~:text=As%20of%202016%2C%20the%20total,are%20the%20number%20of%20cars*.
- *https://www.asirt.org/safe-travel/road-safety-facts/*
- *https://www.nhtsa.gov/*
- *https://wsdot.wa.gov/*