PREDICTING CAR ACCIDENT SEVERITY

Submitted for the Coursera Capstone Project

Abstract

This report analyses factors affecting car accident severity by building a Machine Learning Model.

Submitted by Anna Kourilova

October 2020

Contents

1. Introduction

1.1 Business Problem

According to the WHO, approximately 1.35 million people die each year as a result of road traffic crashes. Road traffic injuries are predicted to become the seventh leading cause of death by 2030. In the USA, The Washington State Department of Transportation Crash Data Portal provides crash information for accidents that occurred statewide. According to the 2019 data, there were 45,524 accidents on all roads. Of those, 235 were fatal crashes, 973 were suspected of serious injury accidents and 2,798 were suspected of minor injury accidents.

To attempt to reduce the frequency of car accidents in a community, a model can be developed to predict the severity of an accident given the current weather, road and visibility conditions, whether the driver was distracted and/or under the influence. The predictive model can be utilized to enhance safety and to help mitigate the likelihood of serious accidents.

1.2 Stakeholders

The algorithm can be useful for the target audience of government officials, drivers and companies in Seattle, Washington. Government officials can forecast the likelihood of collisions and their severity, while drivers can view trends and be more informed of potential risks. As well, companies developing road or car safety technology can analyse these data trends.

Primary Stakeholders:

1. Government Officials (Seattle Public Development Authority)
2. Car Drivers
3. Companies developing technology to improve car/road safety

2. Data Section

2.1 Data Overview and Cleaning

The dataset from 2004-2020 on car accidents that have occurred within the city of Seattle, Washington was used for the project.

The dataset can be found here: https://github.com/annakr-web/Coursera_Capstone/blob/master/Data-Collisions_AK1.zip

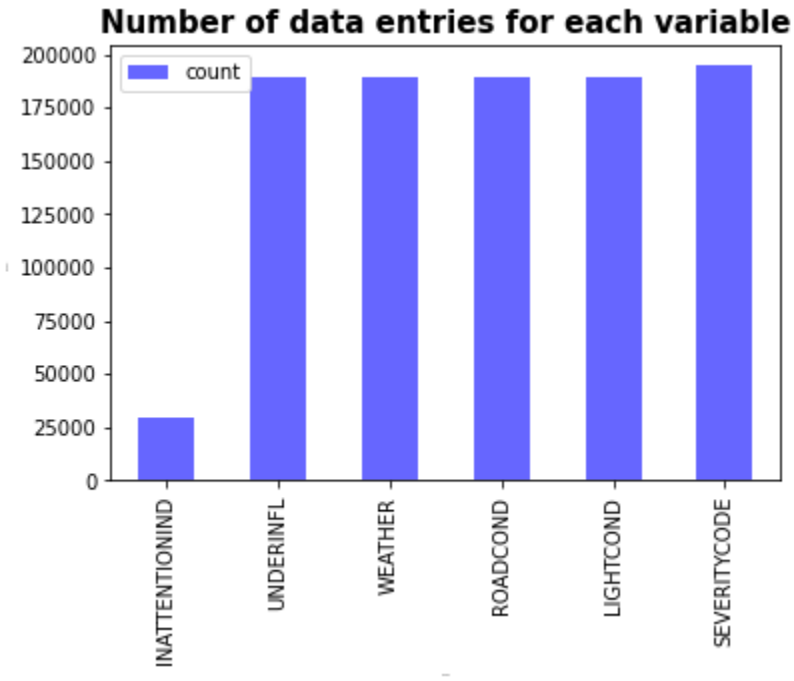
There were numerous challenges with the data set such as variation in frequency of predictor variables and empty values in some columns. Examples of missing values were in columns ‘INATTENTIONID’, ‘SPEEDING’ and ‘PEDROWNOTGRNT’. These columns refer to, respectively, whether the driver was inattentive, whether speeding played a factor in the collision and whether pedestrians were granted right of way. More detail will be provided on the selected features in the next part of the section.

The goal of the developed algorithm was to predict the severity of an accident (categorical variable) by using classification. The predictor variable, ‘SEVERITYCODE’ was initially encoded with values of 1

(Property Damage Only) and 2 (Injury Collision). These were changed to 0 for Property Damage Only and 1 for Injury Collision. For the variables 'INATTENTIONID' and 'UNDERINFL', in the original dataset they were encoded with values of Y, N and no value, which were changed to values of 1 for Y and 0 for N and no value.

For 'LIGHTCOND', Light was given 0 along with Medium as 1 and Dark as 2. For 'ROADCOND', Dry was assigned 0, Mushy assigned 1 and Wet was changed to 2. As for 'WEATHERCOND', 0 was updated for Clear, Overcast was changed to 1, Windy to 2 and Rain and Snow was given a 3. 0 was assigned to the component of each predictor variable which can be the least probable cause of a severe accident. A higher number corresponds to adverse conditions which have a higher probability of leading to a more severe accident. It should also be noted that in order to avoid losing data, variables were codified differently as opposed to deleting the data.

The below chart illustrates the frequency of data for each variable where we can see a much smaller frequency for 'INATTENTIONID' in proportion to the other variables. To solve this problem, arrays were created for each column and encoded according to the initial column data in equal proportions. These arrays were then input into the original columns with Unknown or Other values. The data cleaning process allowed for less data loss and better predictive power for the model.



2.2 Feature Selection

The accident severity was predicted based on five variables. These variables are usually key influences in car accidents. The predictor variable was 'SEVERITYCODE' which measures accident severity as either 1-

property damage collision or 2- injury collision. The criteria used to determine severity were 'WEATHER', which describes the weather at the time of the crash, 'ROADCOND', which describes the road conditions at the time of the collision and 'LIGHTCOND', which describes the light conditions at the time of the collision, 'INATTENTIONIND', which describes whether the driver was distracted, and 'UNDERINFL', which describes whether the driver was under the influence.

3. Methodology

3.1 Data Collection

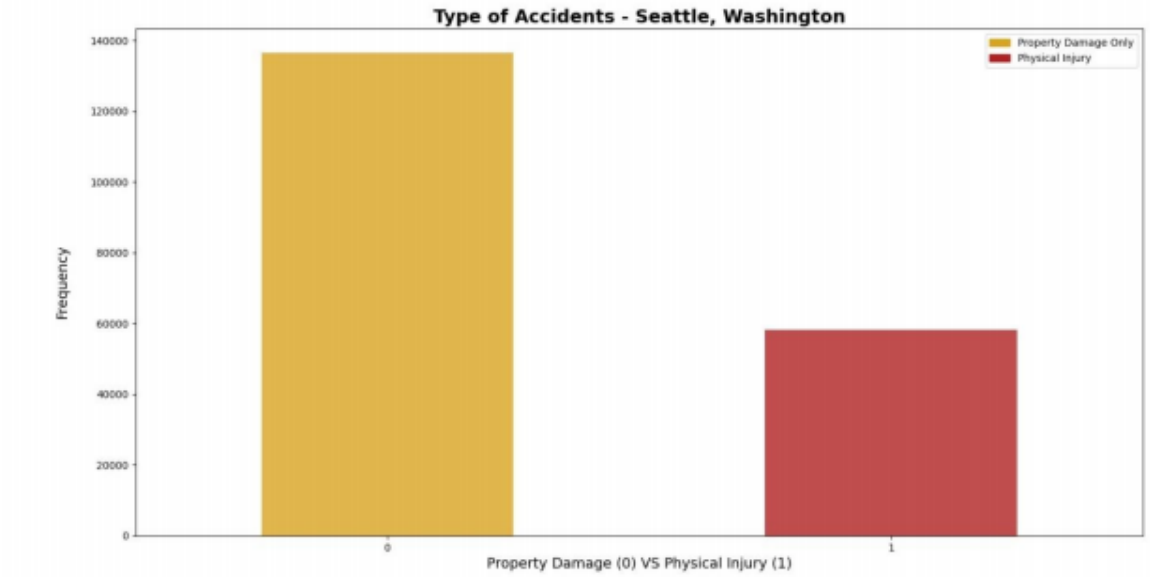
The data used for the project was from collisions that took place in Seattle, Washington between 2004 and 2020. The dataset includes details on collision severity, conditions at the time of the accident and location.

The data can be found on my GitHub page below:
https://github.com/annakrweb/Coursera_Capstone/blob/master/Data-Collisions_AK1.zip

3.2 Data Exploration

After the data cleaning phase, the below representation displays the frequency distribution of each target variable.

This dataset is unbalanced and thus skewed towards Property Damage. When Machine Learning Models are constructed, it is important to have a balanced dataset in order to obtain the most accurate predictive model. To address this issue, SMOTE from the imblearn library was used to create an equally proportioned distribution of the target variables. When training and testing datasets are created, this will act as an unbiased classifier model.



3.3 Machine Learning Model

For this project Logistic Regression and Decision Tree Analysis were used as Machine Learning Models.

First, Logistic Regression is a linear classifier that uses a logistic function based on a combination of features to predict the outcome of a categorical dependent variable based on predictor variables.

Decision Tree Analysis is able to handle high dimensional data with good accuracy. This method partitions on the basis of the attribute values and develops a visualization to facilitate decision making.

These classification methods were selected mainly due to the size of the dataset, considering that the Support Vector Machine model is an inaccurate methodology for large datasets such as this one.

4. Results

4.1 Decision Tree

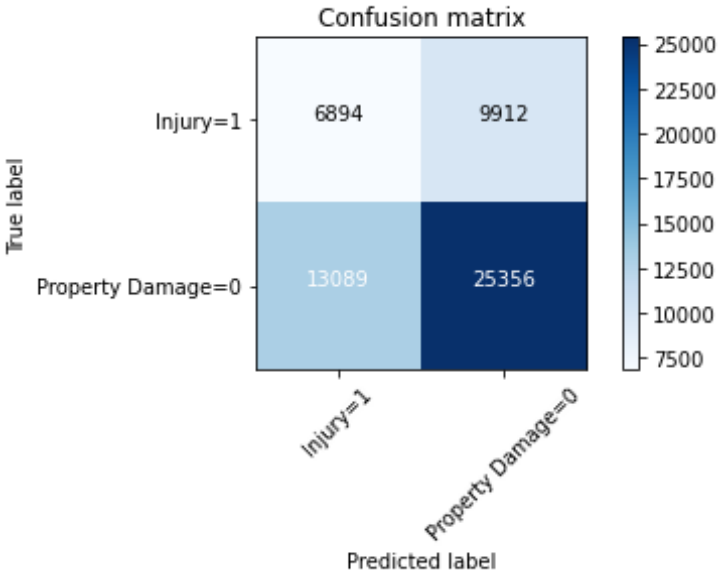
Using the scikit-learn library, a Decision Tree Classifier was used to run the classification on the Car Accident Severity data. Entropy with a max depth of 6 was used as the criteria for the classifier. After applying SMOTE, the new balanced data was utilized to fit and predict the Decision Tree.

The results are summarized below for the Classification Report and Confusion Matrix.

Classification Report

	Precision	Recall	F1 Score
0	0.66	0.72	0.69
1	0.41	0.34	0.37
Accuracy			0.58
Macro Avg	0.53	0.53	0.53
Weighted Avg	0.57	0.58	0.57

Confusion Matrix



4.2 Logistic Regression

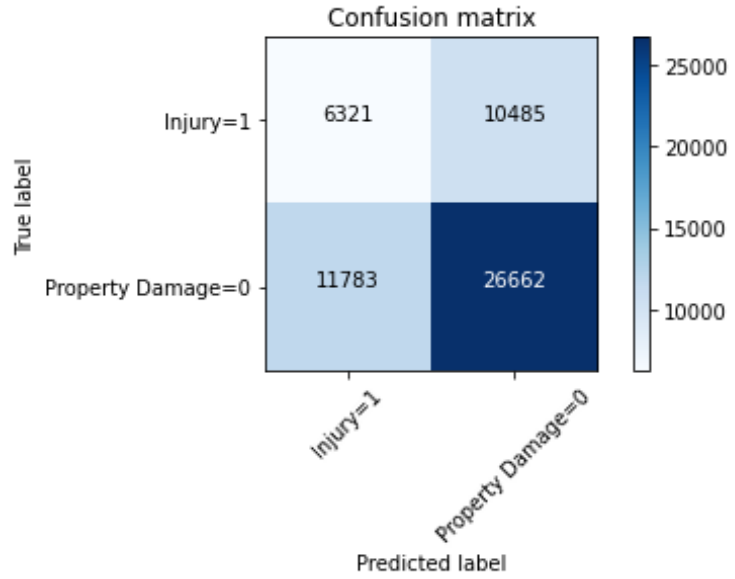
Logistic Regression was used from the scikit-learn library to create the Logistic Regression Classification model on the Car Accident Severity data. The solver used was liblinear and the regularization strength was 0.01. Similarly to the Decision Tree, the post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier.

The results are summarized below for the Classification Report and Confusion Matrix.

Classification Report

	Precision	Recall	F1 Score
0	0.72	0.69	0.71
1	0.35	0.38	0.36
Accuracy			0.60
Macro Avg	0.53	0.53	0.53
Weighted Avg	0.61	0.60	0.60

Confusion Matrix



5. Discussion

Predictive Model	Avg F1 Score	Property Damage (0) vs. Injury (1)	Precision	Recall
Decision Tree	0.53	0	0.66	0.72
		1	0.41	0.34
Logistic Regression	0.54	0	0.72	0.69
		1	0.35	0.38

5.1 F1 Score

The F1 Score denotes the model's accuracy, combining precision and recall. The highest possible value of the F1 Score is 1, which is supported by perfect precision and recall. On the other hand, the lowest value the F1 Score can take is 0, which signifies that either precision or recall is 0.

In the table above, an average F1 score was calculated from Property Damage and Injury. When comparing the two models, we notice that Logistic Regression has a slightly higher F1 score of 0.54, whereas the Decision Tree algorithm has a score of 0.53. While precision and recall of Logistic Regression is a bit superior, the average F1 scores are quite close. It should also be noted that Property Damage has a higher weighting in the model so the F1 score isn't necessarily the best predictor of accuracy.

5.2 Precision

Precision is a percentage measure that explains what proportion of selected variables from the model are material. This is calculated by dividing true positives by true and false positives. Upon looking at the above table, we can assess how accurate the model is at predicting Property Damage and Injury separately. It can be noticed that the highest precision value for Property Damage is Logistic Regression at 0.72 while for Injury it is Decision Tree at 0.41. Overall, the best performing model seems to be the Decision Tree with a precision of 0.66 for Property Damage and 0.41 for Injury. By comparison, Logistic Regression has more of an unbalanced percentage with 0.72 for Property Damage and 0.35 for Injury.

5.3 Recall

Recall expresses, in percentage terms, the total relevant results correctly classified by the Machine Learning algorithm. This is calculated by dividing true positives by true positives and false negatives. The highest recall for Property Damage is the Decision Tree at 0.72 and for Injury it is Logistic Regression at 0.38. The most balanced model for the predictors of the target variable is Logistic Regression with 0.69 for Property Damage and 0.38 for Injury.

6. Conclusion

After doing a comparison of the models in the previous section, we develop a better understanding of the accuracy of the two models individually and overall and their performance for each output of the target variable. The Logistic Regression and Decision Tree algorithms were assessed by their F1 scores, Precision and Recall.

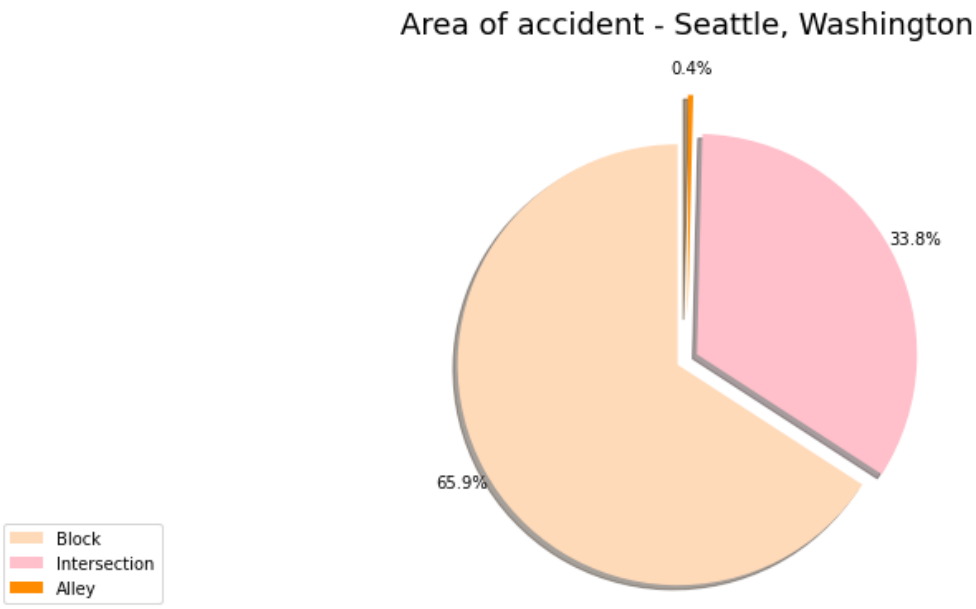
The Decision Tree model has a more balanced distribution of variables for Precision, while Logistic Regression performs better in terms of Recall. Additionally, Logistic Regression has a slightly higher average F1 score of 0.54, only 0.01 higher than the Decision Tree.

It can be concluded that it is advantageous to use these models in conjunction with each other for best results. As well, having a more balanced dataset for the target variable and less blank values for important variables such as ‘UNDERINFL’ (‘SPEEDING’ was excluded for this reason also). With these recommendations, the accuracy and performance of the predictive model could have been improved.

7. Recommendations

After analyzing the data and results of the Machine Learning Models, some recommendations can be made for the relevant stakeholders. The public officials in Seattle can utilize the data to undertake development projects in areas where accidents frequently occur and assess their severity.

The below chart shows a breakdown of accidents in Seattle based on the dataset. The majority of accidents took place on either a block or intersection, hence this should be a helpful indicator of areas where road conditions could be improved or more lighting or signage installed, for example. As well, most accidents occurred under poor lighting, weather and/or visibility conditions.



Both drivers and government officials can make use of the predictive model for accidents occurring based on weather and road conditions. Lastly, these findings and algorithms are beneficial for companies developing technology to improve road safety. These companies can find solutions to minimize future injury and property damage from car accidents based on the predictive algorithm.

8. Sources

https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

http://seattlecollisions.timganter.io/collisions

https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions