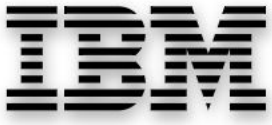

IBM Data Science Professional Certificate

IBM Capstone Project

Predicting Road Accident Severity using Machine Learning

Pamandeep Sangha

October 3, 2020

Abstract

Seattle car accident data were studied. The data was modified and extracted to use in machine learning models. The data attributes used were weather,light conditions, road conditions and accident severity. The models used were KNN, Decision Tree and Logistic Regression algorithms. These algorithms showed equal accuracy of 70% using the Jaccard Index and 57% using the f-1 score. This indicates none of these models were accurate enough to be deployed for future prediction.

Contents

1 Introduction **2**

2 Method **2**

3 Results and Discussion **3**

4 Conclusion **4**

1 Introduction

The problem is that car accidents, or their severity categories, are rarely predicted before they happen. To predict the likelihood of an accident is beneficial to the car driver, passengers, pedestrians, hospitals, police and local residents. Car accident severity is categorised in terms of human fatality, traffic delay, property damage etc, once the accident has already occurred. Various data attributes are also recorded for car accidents such as geographical coordinates, type of road junction, the speed limit, time of day, weather conditions etc. By predicting the accident severity ahead of time, this will likely prevent accidents and engage more caution from drivers. This benefits the stakeholders mentioned such as the vehicle users, the homeowners nearby as roads become blocked off otherwise and traffic builds up if there are road collisions. This in turn costs less for local authorities, courts and public services and ultimately lowers the risk of loss of life. It is proposed that machine learning can be used to accurately predict the accident severity in which drivers in these areas can be notified via their gps, satnav or radio announcements in a timely manner.

2 Method

The methods that were used included modelling from an online dataset, namely car accident collision data from the Seattle Authority in the USA. The data set can be found here:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

To build a good model, the data set should be rich and contain many observations (rows) and various attributes (columns). This should also not include too many attributes such that they become redundant and not benefit the model's accuracy. There should also not be too few attributes such that the model accuracy is low. Therefore, the data attributes that will be used will be severity description, weather, road conditions, and light conditions. It is important to note to only use the data attributes that can be gathered ahead of time to know the accident severity. A lot of the data rows had null values amongst the weather, light and conditions columns so they were dropped. These columns also had to have their categories changed from classifications to numerical values.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight
5	1	Clear	Dry	Daylight
6	1	Raining	Wet	Daylight
7	2	Clear	Dry	Daylight
8	1	Clear	Dry	Daylight
9	2	Clear	Dry	Daylight
10	1	Overcast	Dry	Daylight
11	1	Clear	Dry	Daylight
12	1	Raining	Wet	Dark - Street Lights On
13	1	Raining	Wet	Dark - No Street Lights
14	2	Clear	Dry	Dark - Street Lights On
15	2	Overcast	Dry	Daylight

Figure 1

From there the data were normalised and split into a 70:30 train/test ratio to ensure good levels of out-of-sample accuracy. From there three classifcation models were used: KNN, Decision Tree and Logistic regression. The Jaccard index, f-1 score and Logloss were than calculated.

3 Results and Discussion

The results show that majority of the accidents occurred when the conditions were light and dry. All of the accidents were in severity code 1 or 2, which allowed logistic regression to be used. Severity code 1 outnumbered 2 by a lot which meant oversampling may have caused a skew in results. In all 3 models, the training set accuracy matched the test accuracy at around 70% which is very good. However a testing set accuracy of 70% (Jaccard index) is not that accurate or reliable. The f-1 score showed even lower accuracy at 57% which shows that these models are not accurate enough to be deployed for prediction.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.70	0.57	NA
Decision Tree	0.70	0.57	NA
LogisticRegression	0.70	0.57	0.60

Figure 2

4 Conclusion

In conclusion, the accuracy for these models, while similar, and around 70% were not accurate enough to be used for deployment. In future, I would undersample the severity code 1 data to ensure equal sampling. I would also include the time of day and year of the accidents to see whether this indicates a trend in accident occurrence as well as drivers who were speeding.