Predicting Accident Severity

By Sayam Bhatia

Introduction

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. Accurate predictions of severity can provide crucial information for emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures.

Data

The data chosen for the study consists of all collisions provided by Seattle Police Department (SPD) and recorded by Traffic Records from 2004 to present. This is a link to the dataset https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv . The data has nearly 190,000 entries and the major attributes are as follows:

- Severity code
- Location
- Junction Type
- Collision Type
- Person Count
- Vehicle Count
- Inattention while driving
- Driving under influence
- Speeding
- Weather
- Road Conditions
- Light Conditions

Target data

Of these, the Severity code was taken as the target data as it indicates whether the collision led to human injury or only property damage.

Attributes

Person Count and Vehicle Count were taken as numerical attributes to predict the target data and Speeding was taken as a Boolean.

Junction type, Collison type, Weather, Road conditions and Light conditions were all categorical attributes, but to keep the number of attributes limited to avoid overfitting and delay in processing, Collision type was taken into consideration as it showed highest accuracy post modelling.

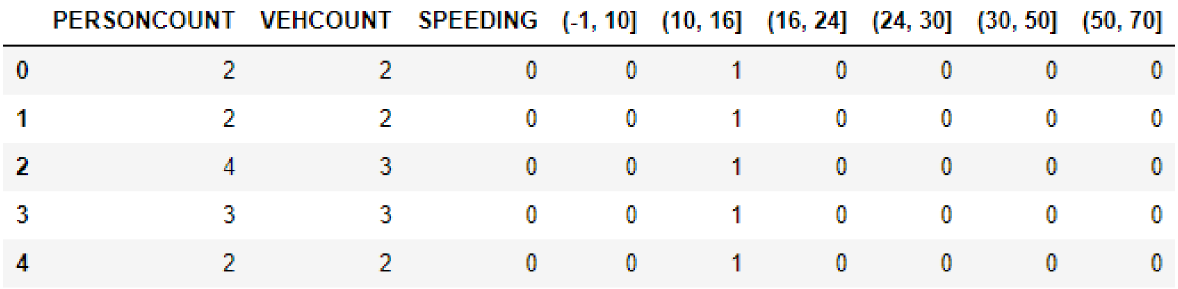
Target	Attribute type	Attribute name
Severity Code	Numerical	Person Count
		Vehicle count
	Boolean	Speeding
	Categorical	Collision type

Data Processing

The Collision type data consisted of code numbers given by the SDOT as per the type of collision. Based on the description of the codes, they were grouped to bins having following description

1. NOT ENOUGH INFORMATION / NOT APPLICABLE
2. MOTOR VEHICLE STRUCK MOTOR VEHICLE
3. MOTOR VEHICLE STRUCK PEDALCYCLIST/PEDESTRIAN
4. MOTOR VEHICLE STRUCK OBJECT
5. DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE
6. PEDALCYCLIST STRUCK MOTOR VEHICLE/ OBJECT/ PEDESTRIAN

The Speeding data was converted from text to numeric and the Collision type data was transposed and appended to the rest of the columns.



	PERSONCOUNT	VEHCOUNT	SPEEDING	(-1, 10]	(10, 16]	(16, 24]	(24, 30]	(30, 50]	(50, 70]
0	2	2	0	0	1	0	0	0	0
1	2	2	0	0	1	0	0	0	0
2	4	3	0	0	1	0	0	0	0
3	3	3	0	0	1	0	0	0	0
4	2	2	0	0	1	0	0	0	0

Data Transformation

The target data was converted to an array and the attributes were fit and transformed to form an array for modelling. The dataset was split into train and test sets, with the following size:

- Train set: (155738, 9) (155738,)
- Test set: (38935, 9) (38935,)

Predictive Modelling

The dataset being dealt with was very large and the data to be predicted was categorical or could even be considered as Boolean as there were only two possible values in the category. Hence, among the data classification models, Decision Tree classifier and Logistic Regression model were chosen for predictive modelling of the dataset.

Decision Tree

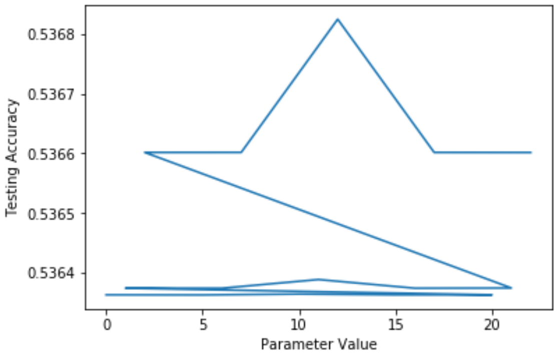
The decision tree classifier was imported from the Scikit learn library. The train set was fit to create a predictive model for a depth range of 1 to 14. The models were used to predict the accident severity for the test set of the data. The accuracy of the models was tabulated for each depth using f1 score and jaccard similarity score. The ideal depth was identified by selecting the depth giving maximum accuracy in both the scores.

Metric	d = 1	d = 2	d = 3	d = 4	d = 5	d = 6	d = 7	d = 8	d = 9	d = 10	d = 11	d = 12	d = 13	d = 14
Jaccard	0.745	0.745	0.752	0.752	0.755	0.754	0.757	0.758	0.758	0.758	0.759	0.758	0.758	0.758
F1	0.674	0.674	0.688	0.688	0.710	0.697	0.710	0.710	0.712	0.712	0.713	0.713	0.713	0.713

As a result, the model with highest accuracy was obtained by fitting the train set at a depth of 11.

Logistic Regression

The Logistic Regression classifier was imported from the Scikit learn library. The train set was fit to create models with varied regularisation values (0.1, 0.01 and 0.001) and solver types (lbfgs, saga, liblinear, newton-cg and sag). The log loss for each combination was determined to find the model with most accurate predictions.



From the above graph, it is observed that the log loss was found to be 0.54 for the entire range of regularization values and different types of solvers with very minor differences.

Model Evaluation

The model predictions were evaluated with the help of Jaccard similarity and F1 scores. The accuracies are listed in the table below.

Algorithm	Jaccard	F1-score	Log Loss
Decision Tree	0.76	0.71	
Logistic Regression	0.75	0.70	0.54

Both the models obtained an accuracy of 75% as per the Jaccard similarity score and an accuracy of 70% as per the F1 score.

Conclusion

In this study, a model was built to predict the severity of an accident while knowing the number of people involved, the number of vehicles involved, whether the vehicles involved were speeding or not and the type of collision that occurred during the accident. The model was able to predict the severity of an accident with an accuracy of 75% with the help of the data set taken for analysis.

The prediction of accident severity is crucial for making better decisions by the emergency response team on the occurrence of such incidents. The prediction of severity can help the response team decide what kind of response would be most suitable for the given conditions.