Predicting the severity of a car accident

Alexandre Rosseto

August, 28, 2020

1.Introduction

1.1 Background

Car accidents are always a big problem in any major city. Being places with lots of traffic, they tend to have a great number of car crashes throughout the year.
This means a lot of money spent in government property being destroyed, people getting injured and, in the worst case, even dying because of these accidents.

Being able to predict the severity of an accident based on some information about it, before reaching the crash site, can improve how hospitals and police departments react to the situation.
Factors like weather, vehicles involved in the accident and light condition, can indicate the probable severity of the crash, and then, hospitals can be better prepared to receive the injured people.
Police and traffic departments can use the same data to reinforce caution in traffic to minimize car accidents and their severity.
This approach can help save a lot of money in government property and, most important, it can help save human lives.

1.2 Objective

This project's objective is to be able to predict the severity of the accidents based on information that can be obtained before any police officer or ambulance reaches the crash site.
By doing this, the people responsible in dealing with the situation can be better prepared to do so.

2. Data acquisition and cleaning

2.1 Data source

The dataset used for this project is the dataset "Collisions—All Years", provided by the Seattle Department of Transportation - Traffic Management Division. This dataset contains records of all types of collisions that happened since 2004 in the city of Seattle.

This dataset contains 37 attributes and 194673 rows. Further information about it can be found at:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

2.2 Data cleaning

The dataset contains a lot of attributes that aggregate no real value to the classification. Id's and other attributes used only to identificate the accident can be discarded in the analysis.

The attributes used in the analysis were: COLLISIONTYPE (Type of the collision), PERSONCOUNT (The total number of people involved in the collision), VEHCOUNT (The total number of vehicles involved in the collision), JUNCTIONTYPE (Category of junction at which collision took place), WEATHER (A description of the weather conditions during the time of the collision), ROADCOND (The condition of the road during the collision), LIGHTCOND (The light conditions during the collision) and SPEEDING (Whether or not speeding was a factor in the collision).

The first problem found in the dataset was on the SPEEDING column: there were only “Y” or null registers. To solve this, all the null values were substituted with “N”, indicating that the driver was not speeding.

The second problem found was the number of null registers on the other columns. The amount for each attribute was counted, and, since they represented a small amount compared to the total of samples on the dataset, all the rows with null values were removed from the dataset, leaving it with 183177 samples.

The next step was to map all the text information contained in the columns to numbers. This was required to be able to use the algorithms of each classifier. For each attribute, the different informations were mapped to an integer, varying from 1 to the number of different informations. For example, in the ROADCOND column all the samples with “Dry” were substituted with 1, all the samples with “Wet” were substituted with 2, and so on.

The dataset after all these transformations is as follows:

	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	SEVERITYCODE
0	2	2	2	2	3	2	1	1	2
1	5	2	2	1	2	2	2	1	1
2	1	4	3	1	3	1	1	1	1
3	4	3	3	1	1	1	1	1	1
4	2	2	2	2	2	2	1	1	2

3. Methodology

The methodology used in this project was to create different models, using different techniques and find the best one amongst them using as parameter the accuracy obtained. The models that were tested were:

- K Nearest Neighbor(KNN)
- Decision Tree Classifier
- Support Vector Machine
- Logistic Regression

It was noticed that the dataset was unbalanced (there were more registers belonging to one class). To minimize this impact, the K-Fold Cross Validation technique was used. This technique divides the dataset in train and test sets K times, using a different subset of the data as train and test samples each time.

4. Modeling and Evaluating

In this section, all the models were created and the training and testing were made using the K-Fold Cross Validation.

The method to calculate the accuracy using this technique is for each iteration of the K-Fold, the models are trained using the training set given by the K-Fold Cross Validation and are tested using the test set given as well. All the accuracies of each model are put together in a list and after all the iterations, the accuracy of each model is given by the mean of all the accuracies calculated at each iteration.

For each model, the overall accuracy was calculated and it is displayed in the following table.

Model	KNN	Decision Tree	SVM	Logistic Regression
Accuracy	71.96%	74.13%	74.77%	71.7%

5. Conclusion and next steps

In this project, a dataset of accidents in the city of Seattle was used to predict the severity of an accident based on some characteristics of the accident.

- The data was treated to optimize the performance of the models created;
- Four different types of models were tested to see which was better;

Based on the results, it's safe to say that the best classifier was the SVM, who achieved the highest accuracy (74,77%).

A possible way to improve the results could be to further analyse the data to find out if using the attributes separately, or in different combinations can improve the results obtained.
The K in the KNN could be optimized, as well as the K in the K-Fold Cross Validation.