Predicting the Severity of a Car Accident

Xiao Peng Yuan

Oct. 10, 2020

1. Introduction

1.1. Background

Car accidents occur every day around the world, a lot of people are involved and get hurt in the accidents. Is there a way to predict the possibility of getting into a car accident and how severe it would be, given the weather and the road conditions, so that people would drive more carefully or even change their travel? This kind of information can be used to prevent the car accidents and reduce the cost of the accidents.

1.2. Problem

Data that might contribute to determining the accident severity might include road condition, weather condition, car speeding, light condition and so on. This project aims to predict the severity of a car accident based on these data.

1.3. Interest

The drivers, insurance companies, road maintenance, government and transport authorities would be very interested in accurate prediction of the car accident severity. Others such as habitants in the area may also be interested.

2. Data acquisition and cleaning

2.1. Data Sources

In this project, the accidents data for Seattle city is used to train the machine learning model, this includes all types of collisions provided by SPD and recorded by Traffic Records, from 2004 to Present.

However, the example dataset has unbalanced labels which will create a biased ML model. Therefore, I downloaded the latest accidents data from Seattle Open Data Portal and use this one instead in my project. The latest dataset is available at http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv?outSR=%7B%22latestWkid%22%3A2926%2C%22wkid%22%3A2926%7D

2.2. Data cleaning

There are several problems with the datasets. First, there were a lot of missing values for the features in the dataset. The missing values for *Road Condition* are filled in with “Unknown”, the same data preparing is applied to the features like *Light*

Condition, *ADDRTYPE*, *Junction Type* and *Weather*. The missing values for *SEVERITYCODE* are filled in with "0".
Second, only the accidents with car speeding factor were marked with "Y" for the feature *Speeding* in the dataset. Mark the rest of the cases with "N".

2.3. Feature selection

After data cleaning, there were 221,525 samples and 37 features in the data. Upon examining the meaning of each feature, it was clear that there were some features that are irrelevant and cannot help to train the model, for example, *OBJECTID*, *INTKEY*, *SDOT_COLCODE*, etc. The information on the features in the dataset can be found at https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.
After examining all the features, I pick up the following features in the dataset to train the machine learning model.

Attribute	Data type, length	Description
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • 3 - fatality • 2b - serious injury • 2 - injury • 1 - prop damage • 0 - unknown
ADDRTYPE	Text, 12	Collision address type: • Alley • Block • Intersection
ROADCOND	Text, 300	The condition of the road during the collision.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
PERSONCOUNT	Double	The total number of people involved in the collision
LIGHTCOND	Text, 300	The light conditions during the collision.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.

3. Predictive Modeling

As explained in the background section, this is a supervised machine learning, the label for the dataset is *SEVERITY*, which describes the fatality of an accident. There are 5 categories in terms of severity:

- 3 - fatality
- 2b - serious injury
- 2 - injury
- 1 - prop damage
- 0 - unknown

Therefore, a classification model is more appropriate in this project to predict the severity of a car accident. In this study, I carried out various algorithms and methods to build the model.

In order to be able to work with different models, the strings in the dataset need to be transformed to numeric input.

3.1. Classification models

Logistic regression, KNN and Decision Trees models were tuned and built. Among the individual models, the Decision Tree model performed the best (75.47% accuracy).

	Logistic regression	KNN	Decision Trees
Accuracy	0.7234	0.7378	0.7547
F1-score	0.6681	0.7006	0.7108
LogLoss	0.6273	N/A	N/A

4. Conclusions

In this study, I analyzed the relationship between the car accident severity and the different factors that might determine the severity. I identified *ROADCOND*, *WEATHER*, *JUNCTIONTYPE*, *SPEEDING*, *PERSONCOUNT*, *LIGHTCOND*, *VEHCOUNT*, *ADDRTYPE* as the most important features that affect the severity of a car accident. I built classification models to predict the severity of a car accident. These models can be very useful in preventing the car accidents and reduce the cost of the accidents.

5. Future directions

In this study, I selected 8 features in the dataset to train the models. Introducing more features such as *INJURIES*, *FATALITIES* may help improve the accuracy, but it will increase the complexity of the models.

Try more different machine learning models is another direction that may improve the accuracy.