Predicting severity of a car accident

Rouslan Gabissov

Augustus 21, 2020

1. Introduction

1.1 Background

Each year many car accidents take place. The consequences of these accidents have different dimensions:

- Human cost: injuries and fatalities of involved persons
- Material cost: damages to vehicles, road infrastructure and other property
- Business cost: cost associated with traffic delays

Everybody recognizes importance of reducing car accidents. Governments are concerned with public health and want to make sure that the number is as low as possible. Businesses want to avoid costs associated with them. And, ordinary people of course want to get to their destination well and on time. Therefore, it is advantageous to everybody to have ability to predict probability and severity of an accident.

1.2 Problem

The features that could predict severity of possible accident might include month, day of week, hour, location, weather condition, vehicle type, roadway configuration, road surface and traffic control. The aim of this project is to predict severity of a possible car accident based on the available data.

1.3 Interest

Car (software in built-in gps systems) and gps navigation software companies (like Google Maps and Waze) could be interested in incorporating predication of severity of possible incidents into their software. In addition, the model could be interesting to public officials interested in improving road safety by for example selecting locations for extra traffic controls or adjusted speed regimes.

2. Data acquisition and cleaning

2.1 Data source

The project's data is related to Canada and provided by Transport Canada and Statistics Canada. The data set can be found on Kaggle here. The set contains 5.86m records observed in the period 1999 to 2014 and 22 columns. All of the features required are present with exception of location. Pre-processing will be required. For instance, there could be duplicated categories in the form of '01' and '1'. The data formatting does not always appear to be consistent. Additionally, some data is unknown or not provided by the jurisdiction.

2.2 Data cleaning

The data set contain many types of indication for missing values: ‘U’, ‘X’, ‘N’, ‘UU’, ‘XX’, ‘NN’, etc. I have set them all to NaN value to make it easier to identify them. To have a clean data set I decided to drop all rows containing at least one NaN value.

The data type of almost all variables was ‘Object’ while all of them were actually integers. Therefore, I had to cast columns’ types to ‘Integer’.

2.3 Feature selection

After examining data it became clear that some of the features are redundant for the purpose of answering business question. For example, there was a feature that was holding year of collision. This feature would not help predict severity of future accidents.
Another example is person level data that contains variables like person sex, age, position, etc. As users of the GPS systems will not input this information before each ride this information has no value for future predictions.

3. Exploratory Data Analysis

3.1 Data set size

After feature selection and data cleaning there were 4.054.430 samples and 12 features in data.

3.2 Target variable

As target variable, I used ‘C_SEV’ which stands for ‘Collision severity’. The variable has two possible values:

- 1: Collision producing at least one fatality
- 2: Collision producing non-fatal injury

3.3 Correlation matrix

Correlation matrix showed that no pair of variables is too strongly correlated; therefore no need to exclude any of them.

4. Predictive modeling

4.1 Model choice

The goal of this project is to classify a possible accident into one of two categories. Thus, the model that will be applied is a classification model.
For the purpose of this project, I decided to use Decision Tree algorithm.

4.2 Model development

I have split data into two sets. One for training the model and another for testing the model.
After training the model, I have predicted the values and compared them to actual values.
The model evaluation showed an exceptional accuracy of 98.4%.

5. Conclusion

In this study I analyzed relationship between severity of an accident and it's collision and vehicle data. I built a classification model to classify whether an accident will include fatalities or not. This model can be very useful for GPS system providers in a number of ways. For example, a GPS software could provide a warning asking to be careful because there is a chance of an accident with fatal end, or a GPS software could calculate a road taking into account a chance of an accident (not part of this project) and severity of it.