# Predicting the Severity of Car Accidents

Dimitris Manolakis

October 13, 2020

## 2. Data acquisition and cleaning

### 2.1 Data sources

The data used on the project come from the UK Car Accidents dataset which is collected from the UK Department for Transport. The dataset starts at 2005 and ends at 2015 and is composed by three different csv files (Accidents, Causalities, and Vehicles). It was selected because it offered a great amount of variables (features) for each accident like personal information on the driver, environmental conditions and vehicle information. Also, the accidents were classified into three types of severity (1. Fatal, 2. Serious, 3. Slight) which gave promise for a model with more specific predictions than most datasets that offered binary classification for the accidents severity. The dataset can be found on Kaggle.

### 2.2 Data cleaning

The 3 different csv files were joined into one data frame using the common column 'Accident Index'. The data was in a good condition as most of the variables were already categorical variables (integers) and there were very few null values (172 null values for the 'Time' column only).

Multiple columns such as ‘Weather Conditions’, ‘Road Surface Conditions’ and others included Unknown or Other values marked as -1 or 9 accordingly. All the rows that had either value was removed from the data.

Additionally, the ‘Age Band of Driver’ column included values for extremely under aged drivers that were under 15 years old. The rows containing these values were removed from the data.

Finally, a number of functions were created that converted categorical variables to the corresponding string that described the variable’s value (e.g. for ‘Day’ column: 1 converted to ‘Monday’, 2 converted to ‘Tuesday’ etc.). These functions were needed in order to better explore the relations between the data and create easy to read statistical plots.

## 2.3 Feature selection

In total, 11 features were selected in order to be further analyzed as predictors for the model. The following table contains these features:

Table 1: The 11 features selected from the dataset

| | |
|---|---|
| Features | Day of Week, Time, Road Type, Light Conditions, Weather Conditions, Road Surface Conditions, Urban or Rural Area, Vehicle Type, Sex of Driver, Age Band of Driver, Age of Vehicle |

The criterion I used to select these features was based on finding the factors that were determined before the accident happened. For example, I did not want to use information on the amount of vehicles that ended up in the accident or personal information on the casualties. Instead I wanted information that can be known before the accident happens like driver’s age or the vehicle’s type.

Analyzing the relation between each one of the features with the Accident Severity was a difficult task for 2 reasons. Firstly, I needed a statistic that measures correlation between categorical variables. After

some research on the topic, I found Cramer's V. It is based on a nominal variation of Pearson's Chi-Square Test, and comes built-in with some great benefits:

1. Similarly to correlation, the output is in the range of [0,1], where 0 means no association and 1 is full association. (Unlike correlation, there are no negative values, as there's no such thing as a negative association. Either there is, or there isn't)

2. Like correlation, Cramer's V is symmetrical

After testing the association between the variables and the severity, I found them to be highly unassociated. And this takes us to the second reason that analyzing the correlation between the severity and the features is a difficult task. The 84% of the accident have a Slight severity which makes the dataset very imbalanced since this is the feature we try to predict. The solution to this problem, at least for finding associations between the variables and the accidents was to test them against the total amount of accidents each category was involved. For example, 83% of the accidents happened during Fine weather which makes the feature not a good predictor for the model. In some cases I further explored the frequency of Fatal or Serious accidents against a particular feature in order to gain deeper insights. The following table contains the features which was selected as predictors for the model.

Table 2: The 4 features selected as predictors

| | |
|---|---|
| Features | Light Conditions, Urban or Rural Area, Age Band of Driver, Hour |