Predicting Crash Severity and Fatalities for NZ Road Accidents

Domain Background

Car crashes and road accidents could be considered an old topic. Yet, with the progress of the technology involved and the capabilities of these sophisticated machines, it is ever more important to have tools and means available to mitigate their occurrences, as well as their implications and consequences for the people involved. Thanks to the advancement of technological and analytical tools in the last two decades we are now able to better understand how crashes happen. This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

Many attempts have been taken by many professionals, scholars and government agencies to provide produce these models; each with different goals, ways of measuring success and precision of their results. The *Benchmark* section lists a few research papers –some of which are being used in production systems– that offer a myriad of different approaches to the problem at hand.

On a personal note, I'm looking to move to New Zealand. I was drawn to this problem while looking for an interesting dataset and problem that I could use to build a portfolio; which I then intend to use for networking purposes with professionals from NZ in my search for job opportunities.

Problem Statement

Predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on the data available and how well the system or model has been defined. However, if the dataset's features are clearly defined and if there's a thorough description of how this data is collected we have much better chances of arriving at a usable model. In the dataset I'll use data associated with car crashes come in a hybrid mode; meaning we have both categorical and numerical features. This allows us to treat the problem from a mathematical approach and to use performance metrics such as R^2, *ROC curves*, *precision*, *accuracy* and *F Scores*.

Dataset & Inputs

The dataset that we will use comes directly from the *New Zealand Transport Agency* and is made available through the website *data.govt.nz;* a repository of open datasets related to all kind of activities throughout New Zealand and published by the central government. It contains data from car crashed since January 1st of the year 2000 until the present day and it's automatically updated on a quarterly basis.

The dataset (its CSV version including 2018Q2) can be found in its corresponding landing page here, or can be directly download here. In addition, the NZ Transport Agency has made available a definition for each feature included in the data here.

It's important to note, as the description states, that: "not all crashes are reported to the NZ Police. The level of reporting increases with the severity of the crash. Due to the nature of non-fatal crashes it is believed that these are under-reported".

Solution Statement

Given that we have a rich dataset of 655,698 training examples, each with 88 features; and that this data has both a good history going back to the year 2000 and also a good update policy; we have a great opportunity to produce a viable predictive model that could be used by emergency services all around NZ to improve the response time and also to produce a proper response depending on the severity of the car crash.

Moreover, given all the features that are available, we have also an opportunity to try to uncover hidden patterns and structure in the data that could be leveraged to better understand which factors are more indicative or play a bigger role in such accidents. Thus, providing valuable insights towards preventing them from happening in the first place (such as better signaling, road maintenance, or even improve road construction planning).

Also, using some dimensionality reduction technique such as *PCA*, *Random Projection* or *ICA* could provide us with new and unseen features while also simplifying dataset used to train the classifiers.

Benchmark Model

After some research, I've been able to find a few (non-free) research papers describing different models developed precisely for this problem. However, while all of them together are a confirmation that what we attempt to do is possible, I was only able to find one which provides (in its abstract) some performance metrics we can use as a benchmark.

These papers are:

1. Prediction of Road Accident Severity Using the Ordered Probit Model
   2014 - RuiGarrido[i] AnaBastos[i] Anade Almeida[i] José PauloElvas

2. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice
   1996 – C.J. O'Donnell; D.H. Connor

3. Predicting Severity and Duration of Road Traffic Accident
   2013 - Fang Zong; Huiyong Zhang; Hongguo Xu; Xiumei Zhu; and Lu Wang

4. Severity Prediction of Traffic Accident Using an Artificial Neural Network
   2016 - Sharaf Alkheder; Madhar Taamneh; Salah Taamneh

As Paper [4] mentions in its abstract:

"The overall model prediction performance for the training and testing data were 81.6% and 74.6%"

And so, this will be our benchmark, an ANN with an overall accuracy of at 74.6%.

Evaluation Metrics

Considering we will develop 3 different models, each chained to the previous one, we must have good metrics at each step.

For the dimensionality reduction part, we don't really have a metric we can use. However, we will perform a thorough analysis of its results to assess whether we can use it for the other steps in our ML Pipeline.

Next, for the unsupervised part where we will attempt to uncover unseen structure and patters hidden in the data we will try different unsupervised algorithms (*Hierarchical Clustering*, *DBSCAN*, *GMM*) and measure their performance with the *silhouette coefficient* when possible.

For the *Crash Severity Multi-Class Classifier,* we will use an F score and discuss what's the optimal betta value to use.

Finally, for the *Fatalities prediction* part we will explore to paths. A regression and a classification. For the regression approach we will use R^2 as a performance metric. While for the classification approach we will again use an F score with different betta values.

Project Design

Here's a summarized workflow taken from all that has been said thus far:

1. Extensive EDA: explore structure, values and variance in the data. Identify features that could be disregarded due to redundancy, high correlation or very low variance.

2. Implement Dimensionality Reduction techniques in an attempt to engineer some additional features that may be of value.

3. Implement unsupervised learning algorithms to uncover patterns and structure that could leveraged for the final classifier.

4. Train a classifier to predict crash severity.

5. Train a regressor/classifier to predict number of fatalities when the crash class is *Fatal*.