# Actuaries Institute Kaggle Competition 2016 - VicRoads

## Data

The competition uses data donated by the Victoria Roads Corporation (VicRoads). The accident data initially included all vehicle accidents attended by the Police in Victoria. This data set is already in the public domain. The Road data was collected by an organisation called iRAP (irap.org) by driving a vehicle on all A,B and C designated roads in Victoria. The vehicle took measurements and photographs approximately every 100m along the road. The data was then derived from these measurements and photographs.

Joining the accident data to the road information was not an easy exercise. Accidents which were not on A,B or C roads were eliminated, and the Police recorded location needed to be mapped as Latitude and Longitude. These were then joined to a road location taking into account proximity and direction of travel. It is likely that some errors have been made. In addition many accidents occur at intersections and these have been attributed to only one of the intersecting roads.

The road data is assumed to be static over the whole period. In practice it corresponds to the more recent accident data and the older accident history may be less relevant to fitting a model than the more recent history.

The competition requires the prediction of accident cost on a particular road in a particular calendar quarter.

## Separation into Training and Test Sets

Accidents on a 100m section of road are relatively rare in a quarter. The purpose of the competition is to find valuable predictive factors in the nature of the road infrastructure and surroundings. For this reason we wish to exclude nearest neighbour type models that would predict accidents based on the neighbouring section of road without solving what it is about the road that influences the accident rate (eg traffic volume, weather, speed limit, curvature, safety barriers). In addition we want to predict future accidents, so would ideally want to time-slice the last period of data as a test-set to ensure that the winning model is a good predictor of out-of-period accidents. Models that predict future accidents based solely on past accidents on the same section of road would not be useful for judging whether improving the road infrastructure would reduce accidents.

To penalise nearest neighbour models and historic accident models the test data was chosen as follows:

* The roads were chopped into sections (called BLOCKs) around 20km long. These sections are either all out or all in the test data, so the nearest neighbour of a test row may be 10km away.
* Roughly half the test data is from BLOCKS where all periods for the block are in the test data, so there are no historic accidents for these blocks in the training data.
* The remaining half of the test data is the last year (4 quarters) of accidents (from April 2015 to If you are using cross validation to choose the best models you may want to duplicate this approach in creating the cross validation sets. Alternatively if your model only uses road infrastructure as predictive variables than you can randomly allocate the training rows into cross validation subsets. A code sample in R shows how you could use BLOCKs to create a validation subset.

## The Cost function

The Cost function should not be seen as “How much anyone thinks a life is worth”. It is merely an algorithm to prioritise road infrastructure. The highest monetary cost is associated with severe injuries rather than fatalities, but avoiding fatalities is more significant in prioritising infrastructure. The minor accidents are included as a cost not because they are significant in themselves, but are likely to have predictive value for more severe accidents.

## Data Fields

The data fields are described in the Excel file Data Dictionary.xlsx

## Other Data

You may make use of other publically available information to improve your model with the exception of Victorian accident statistics or hospital admissions. Examples of the type of information that may be relevant are:

* Weather Events and Sporting Events
* School holidays and Long Weekends
* Road Construction Schedules and temporary speed limits
* Vehicle Registrations is different neighbourhoods
* Information or processes to attribute roads to neighbourhoods